Data Mining Report
Analysis of Global Education Enrolments

Mohd Umar Yugal Verma Adam Bielecki
Verna Perumal Kormaz Deniz Labidi Samar

November 23, 2025

Abstract

This report documents the solution for the Data Mining Sheet 4 exercises. The objective was to gather, inspect, clean, and visualise real-world datasets regarding Global Education. The analysis covers three distinct levels of education: **Primary (Ex 1)**, **Secondary (Ex 2)**, and **Tertiary (Ex 3)**. Across all datasets, consistent data cleaning strategies were applied to handle non-numerical units and missing gender attributes. The results highlight distinct trends, particularly the shifting gender balance as education levels increase.

1 Exercise 1: Primary Education Analysis

1.1 Introduction & Data Cleaning

The first dataset contained annual primary education enrolment figures. Key data quality challenges included mixed units (percentages vs numbers) and missing male data. The cleaning process involved:

1. **Renaming:** Simplifying *Reference Area* to *Country*.
2. **Filtering:** Removing rows where *Units of measurement* was not "Number".
3. **Feature Engineering:** Calculating Male enrolments using: $Male = AllGenders - Female$.

1.2 Visualisation (Primary Education)

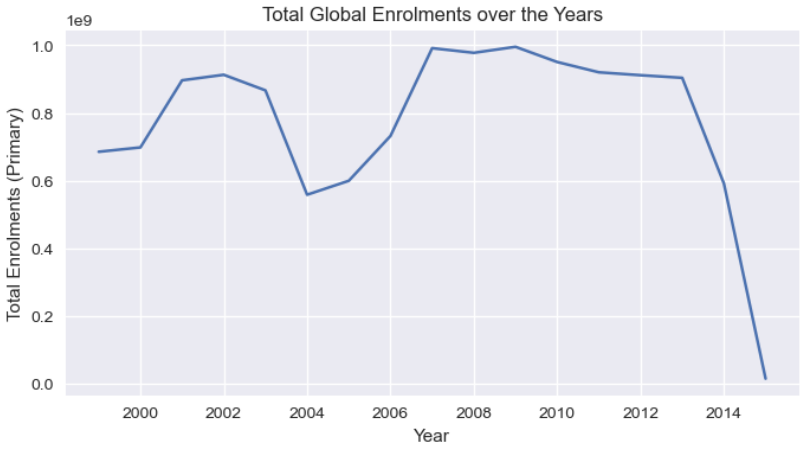

Figure 1: Ex 1: Total Global Primary Enrolments over the years.

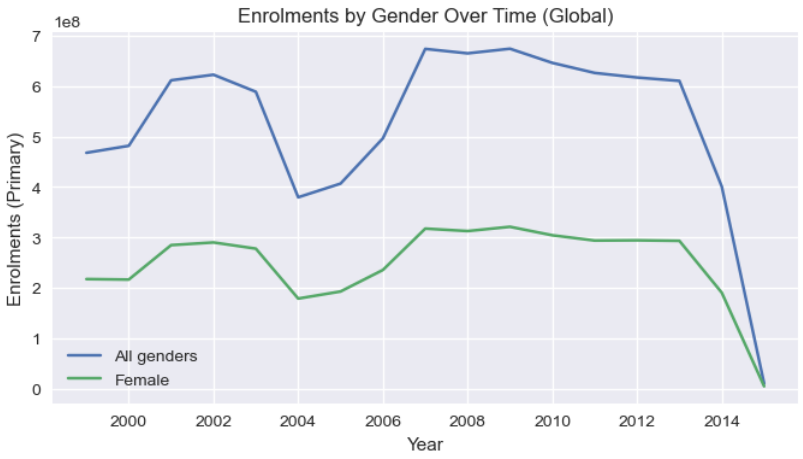

Figure 2: Ex 1: Primary Enrolments by Gender (All Genders vs Female).

Figure 3: Ex 1: Derived Male vs Female Primary Enrolments.

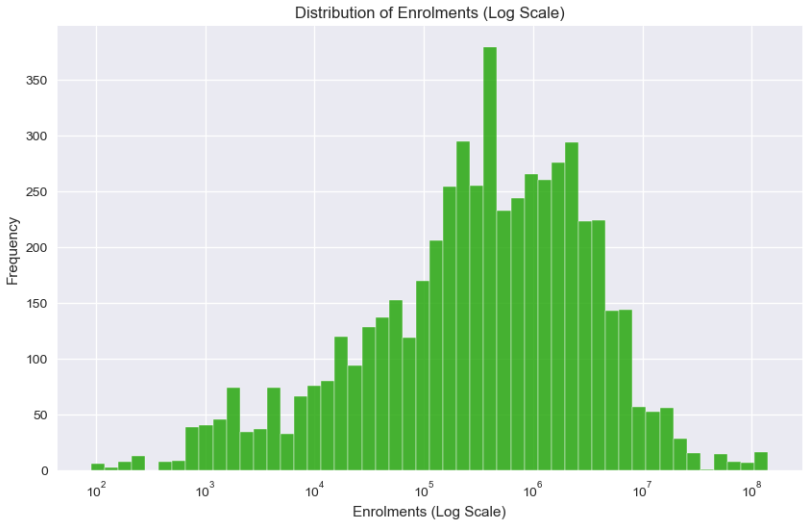

Figure 4: Ex 1: Distribution of Primary Enrolments (Log Scale).

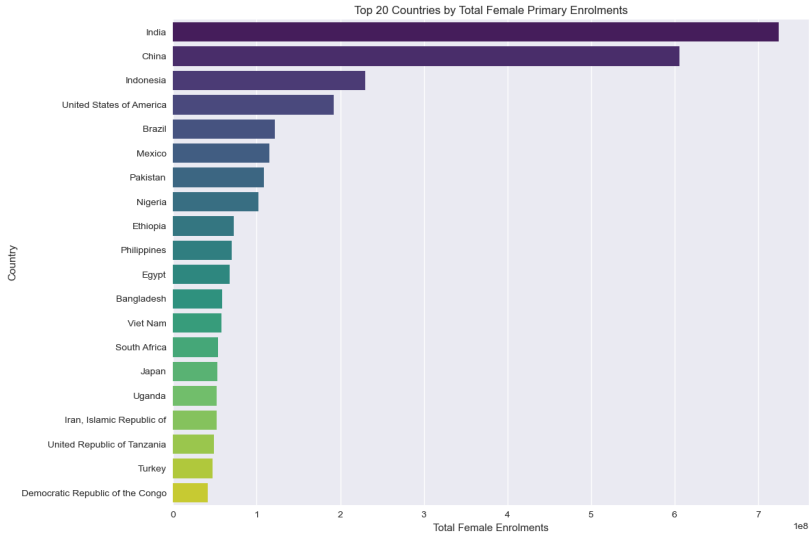

Figure 5: Ex 1: Top 20 Countries by Total Female Primary Enrolments.

2 Exercise 2: Secondary Education Analysis

2.1 Overview

Following the methodology established in Exercise 1, we analyzed the `secondary.csv` dataset. The cleaning steps were identical, focusing on unit standardization and deriving the male population count.

2.2 Visualisation (Secondary Education)

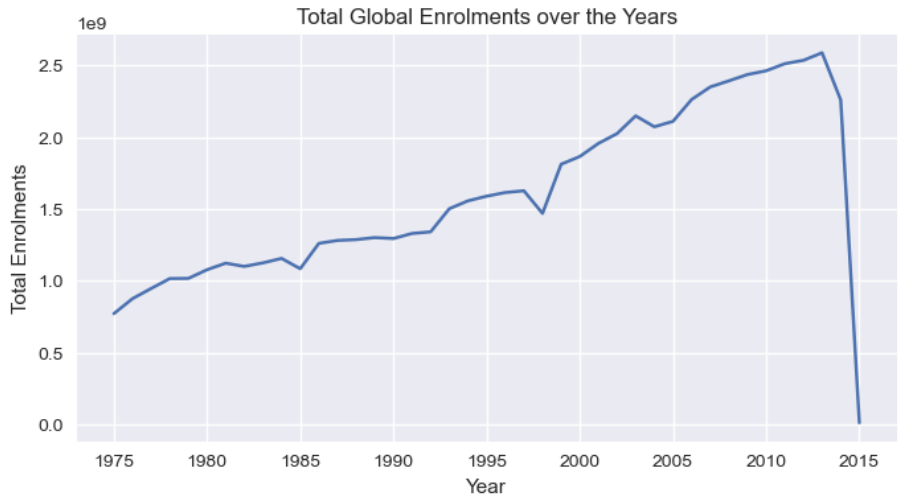

Figure 6: Ex 2: Total Global Secondary Enrolments. A steady increase is observed until the reporting cut-off.

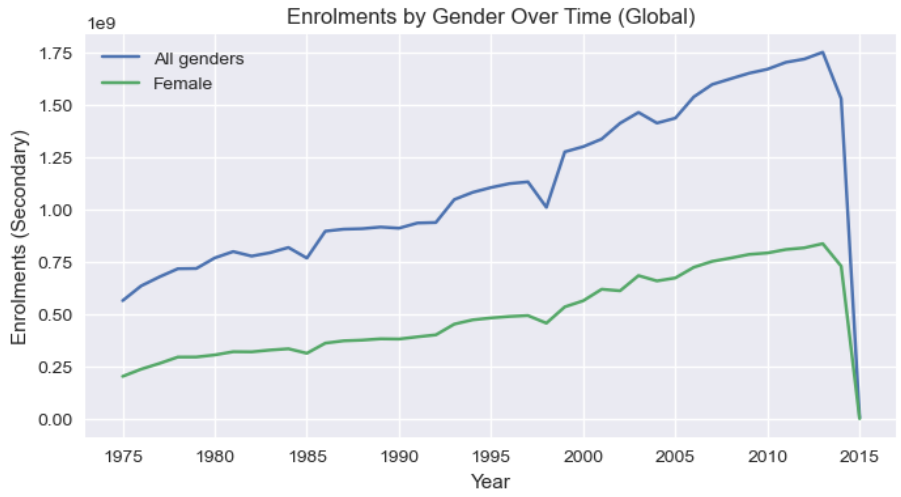

Figure 7: Ex 2: Secondary Enrolments by Gender (All Genders vs Female).

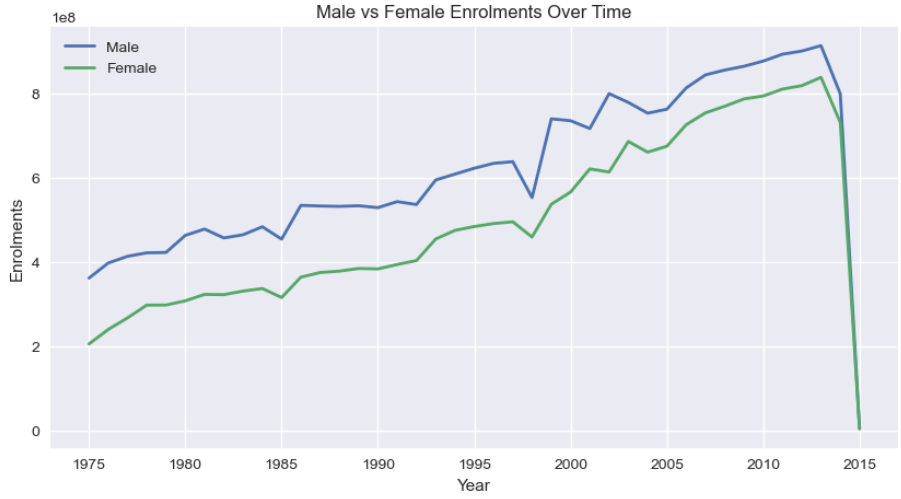

Figure 8: Ex 2: Male vs Female Secondary Enrolments. Males consistently outnumber females, though the gap appears narrower than in primary education.

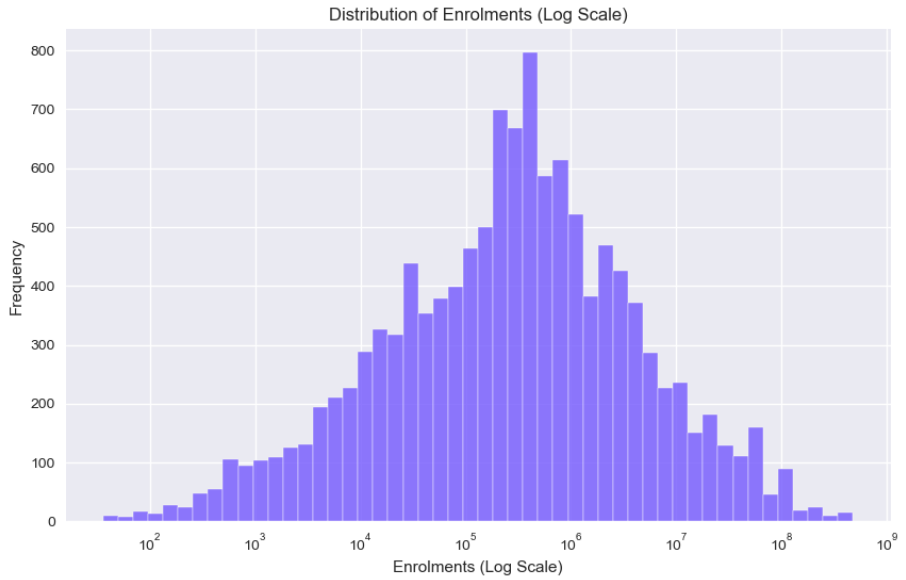

Figure 9: Ex 2: Distribution of Secondary Enrolments (Log Scale).

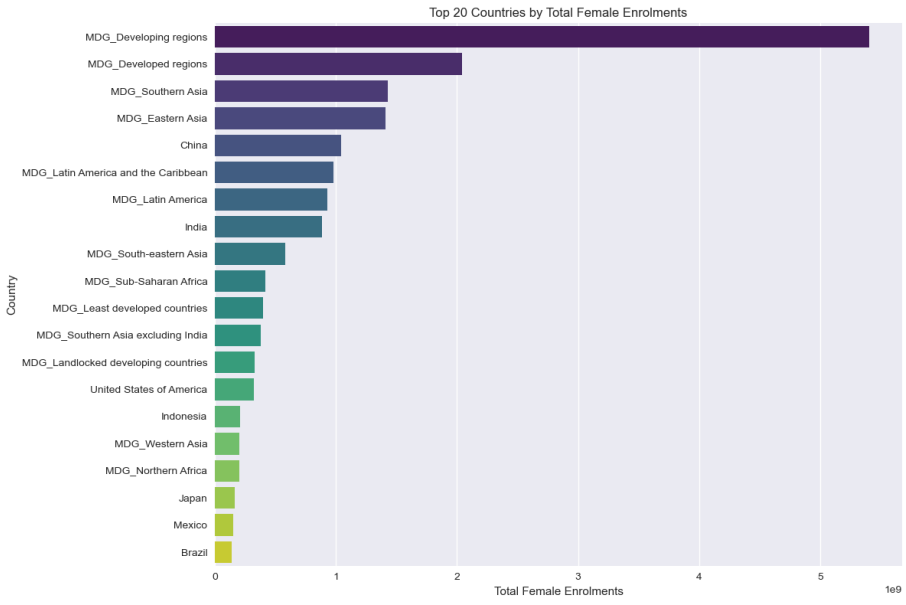

Figure 10: Ex 2: Top 20 Countries by Total Female Secondary Enrolments.

3 Exercise 3: Tertiary Education Analysis

3.1 Overview

The final analysis focused on `tertiary.csv`, representing higher education enrolments. While the data structure mirrored the previous datasets, the trends revealed in this specific education level differed significantly from Primary and Secondary education.

3.2 Strategy and Processing

The cleaning strategy remained consistent to ensure comparability:

- **Cleaning:** Rows with non-numerical units were removed. Redundant columns (*Age group*) were dropped.
- **Transformation:** The "Male" attribute was derived via the formula $Male = Total - Female$.
- **Observation:** Unlike the previous datasets, this file appears to contain aggregated regional data (e.g., "Developing Regions") alongside individual country data, as seen in the Top 20 plot.

3.3 Visualisation (Tertiary Education)

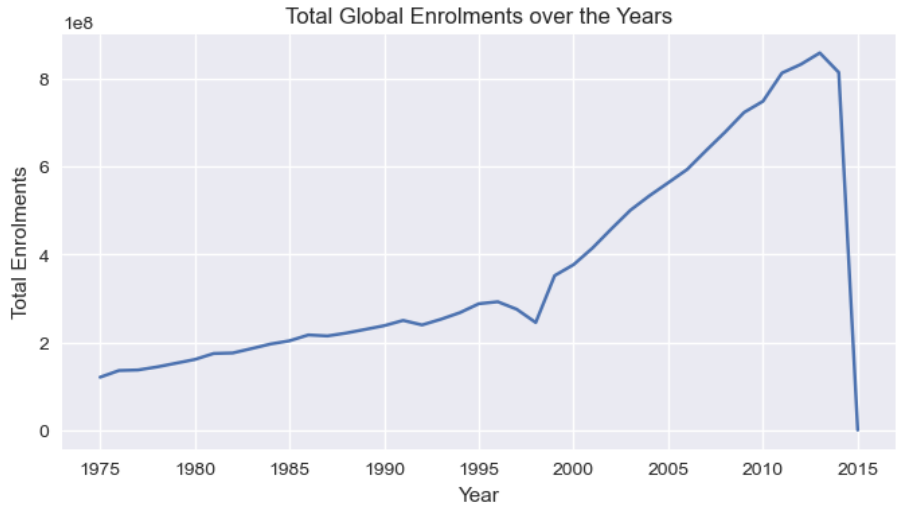

Figure 11: Ex 3: Total Global Tertiary Enrolments over the years. The growth curve is steeper compared to primary education, reflecting the global expansion of higher education access.

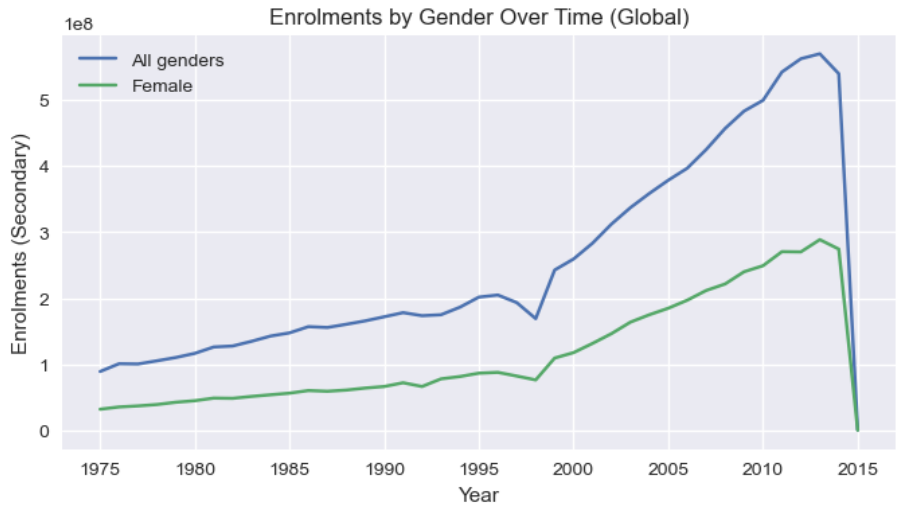

Figure 12: Ex 3: Enrolments by Gender (All Genders vs Female). The female proportion appears visibly closer to the total compared to previous exercises.

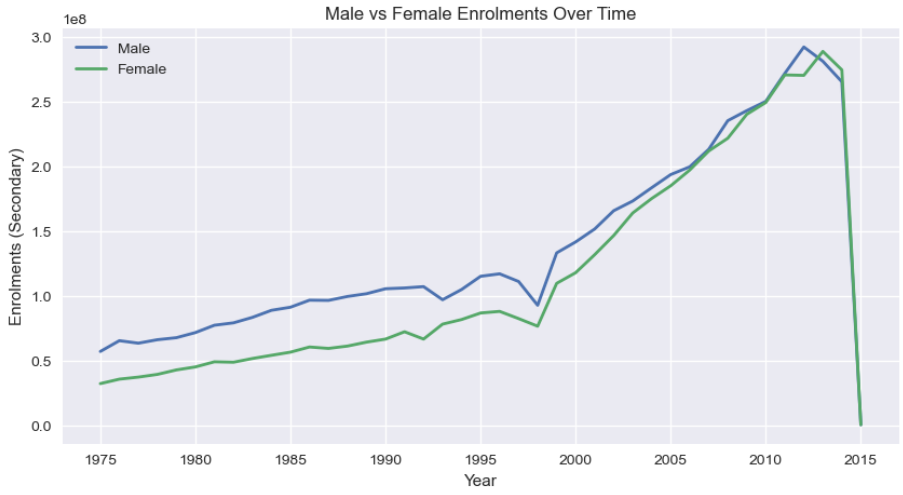

Figure 13: Ex 3: Male vs Female Tertiary Enrolments. **Significant Finding:** Unlike primary and secondary education, female enrolment in tertiary education catches up to and eventually surpasses male enrolment around 2005.

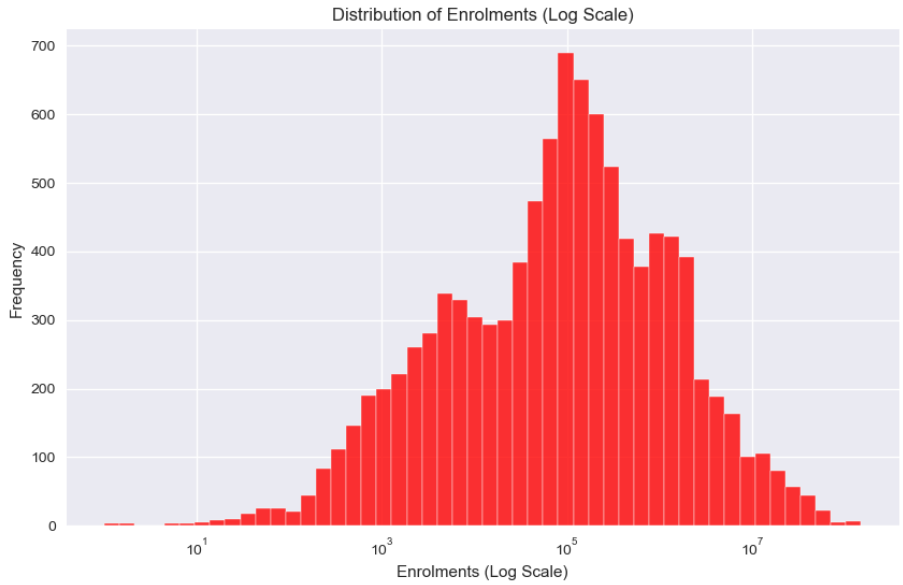

Figure 14: Ex 3: Distribution of Tertiary Enrolments (Log Scale). The distribution remains right-skewed.

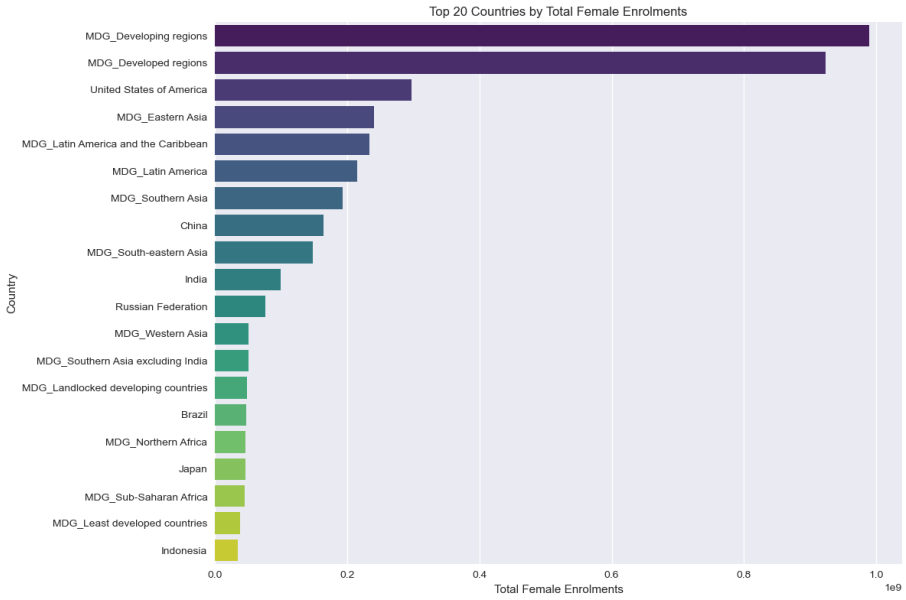

Figure 15: Ex 3: Top 20 Regions/Countries by Female Enrolment. Note the presence of "MDG_Developing regions", indicating that this dataset contains aggregated regional data.

4 Conclusion

The comprehensive analysis of Primary, Secondary, and Tertiary education reveals a clear narrative. While Primary and Secondary education show a historical male dominance in enrolment numbers, **Tertiary education shows a reversal of this trend**, with female participation exceeding male participation in the most recent years of the dataset.

Common data quality issues—specifically the lack of explicit male data and inconsistent units—were successfully mitigated through a unified Python preprocessing pipeline, allowing for accurate comparative analysis across all three education levels.