RWANDA POLYTECHNIC

HUYE COLLEGE

MODULE NAME: DATA MINING AND WAREHOUSING.

MODULE CODE: ITLDM801

DEPARTMENT: ICT

OPTION: IT

LEVEL: 8 YEAR 4

CLASS: IT Btech

Date: 12th Feb, 2026

MULTI-CAMPUS EDUCATION DATA PREPROCESSING PIPELINE

GROUP 3: MEMBERS

NAMES	**REG NO**	**MARKS /100**
CYUMUGISHA Dorine	25RP19746	
IZABAYO Clementine	25RP21648	
RUHAMO Rose	25RP21044	

Table of Contents

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

This report presents the design, implementation, and technical specifications of a Multi-Campus Education Data Preprocessing Pipeline. The pipeline integrates student, course, and assessment datasets from multiple campuses to create a clean, unified, gold-level dataset ready for analytics, reporting, and machine learning.

2.1 BACKGROUND

Educational institutions often operate multiple campuses, each maintaining independent student, course, and assessment records. Variations in data formats, missing values, duplicates, and inconsistent codes make cross-campus analysis challenging.

2.2 Problem Statement

The datasets contained several data quality issues, including missing values (Gender, DOB, Marks), duplicate student and assessment records, outliers such as invalid marks or credits, inconsistent formatting of dates and codes, and noisy textual entries. These problems prevented reliable analysis and required systematic preprocessing.

3. PROJECT OVERVIEW

This project:

- ✓ Loads and cleans all raw data using Python and Pandas.
- ✓ Standardizes formats, imputes missing values, removes duplicates, and corrects outliers.
- ✓ Integrates datasets across campuses into a gold-level dataset.
- ✓ Performs feature engineering to enhance analysis readiness.

3.1 Data Sources

Data Source	**Format**	**Description**
Students	CSV	Student profiles including Gender, DOB, Program, Level, Intake_Year
Courses	CSV	Course code, Course Title, Credits
Assessments	CSV	Marks, Attendance Rate, Assessment Type, Academic Year, Semester

Table 1: Data Sources

3.2 Project Objectives

The objectives of this project are to:

- ✓ Standardize and consolidate campus datasets.
- ✓ Remove duplicates, outliers, and inconsistencies.
- ✓ Produce a gold-level dataset for analytics.
- ✓ Engineer features for academic performance and risk assessment.
- ✓ Ensure the dataset is analysis-ready with no missing critical values.

4. ISSUES FOUND AND EVIDENCE

- ❖ **Missing Values**

Missing values were visualized using bar plots for easy identification of problematic columns.

Observed in multiple datasets:

- ✓ **Students:** Gender and DOB

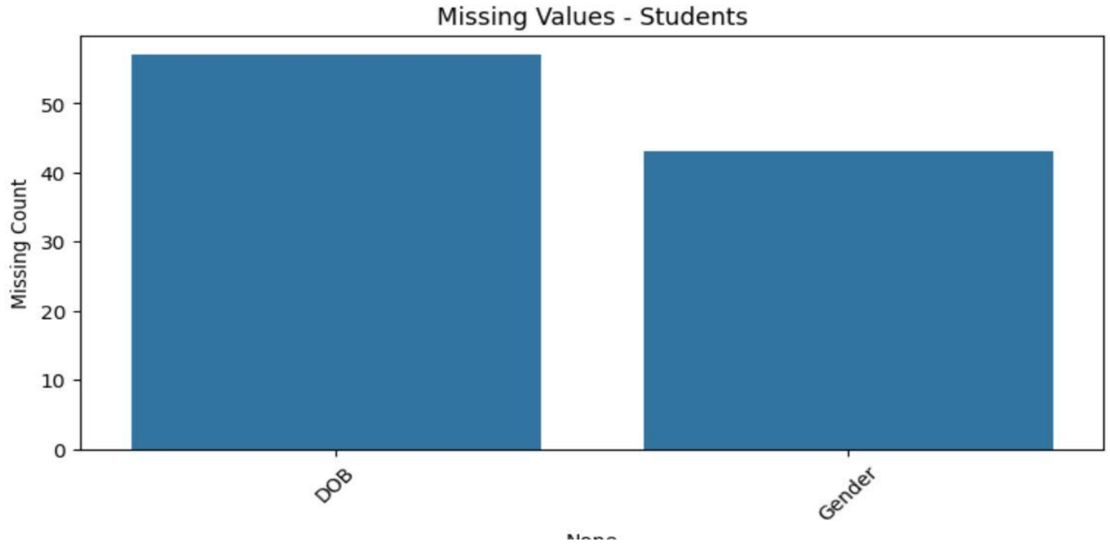


Figure 1: Missing values in students

- ✓ **Assessments:** Mark and Attendance_Rate

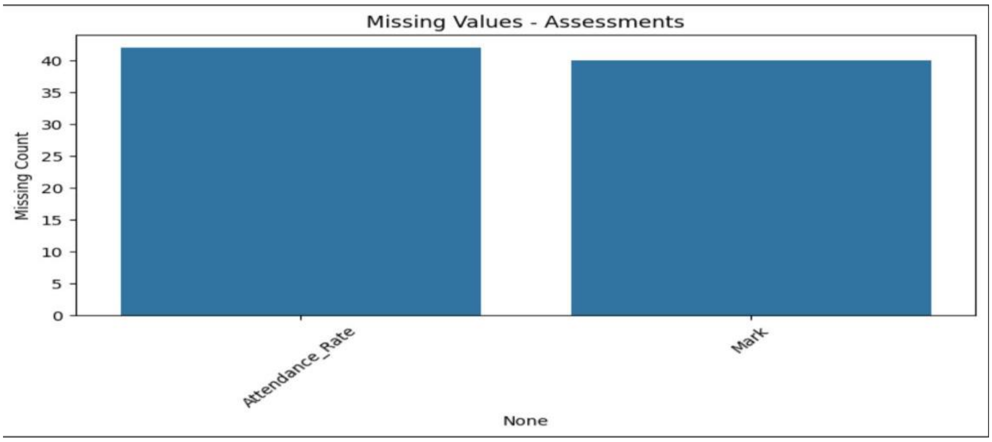


Figure 2: Missing values in assessment

There were no missing values in courses.

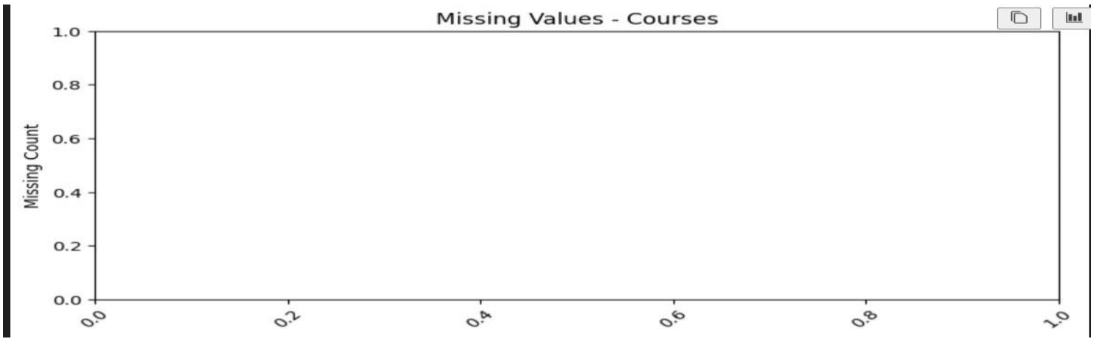


Figure 3:Missing values in courses

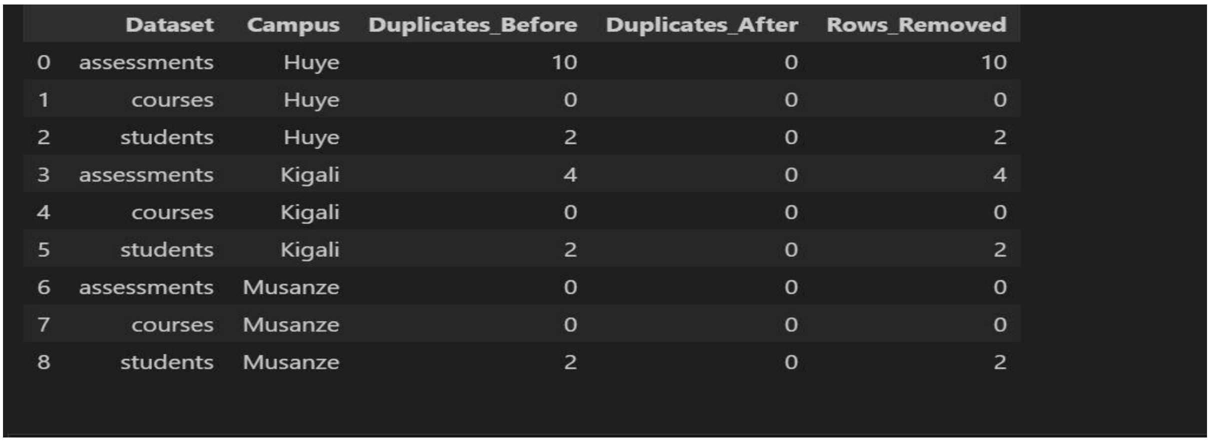
❖ Duplicates Assessment

Duplicate rows were detected per dataset:

✓ Students: duplicates based on Student_ID

✓ Assessments: duplicates based on Student_ID, Course_Code, Academic_Year, Semester

✓ Courses dataset: no duplicates detected



	Dataset	Campus	Duplicates_Before	Duplicates_After	Rows_Removed
0	assessments	Huye	10	0	10
1	courses	Huye	0	0	0
2	students	Huye	2	0	2
3	assessments	Kigali	4	0	4
4	courses	Kigali	0	0	0
5	students	Kigali	2	0	2
6	assessments	Musanze	0	0	0
7	courses	Musanze	0	0	0
8	students	Musanze	2	0	2

- ❖ **Outliers, Noisy Data and Inconsistent Data**

Data quality issues were noted in numeric fields (e.g., Mark outside 0–100, negative Credits) and text fields (extra spaces, inconsistent capitalization).

- ✓ Marks outside valid range (e.g., -5, 105)
- ✓ Attendance > 1 or negative
- ✓ Credits negative or zero
- ✓ Dates in different formats (Assessment_Date, DOB)
- ✓ Text inconsistencies: extra spaces, inconsistent capitalization in Program and Full_Name

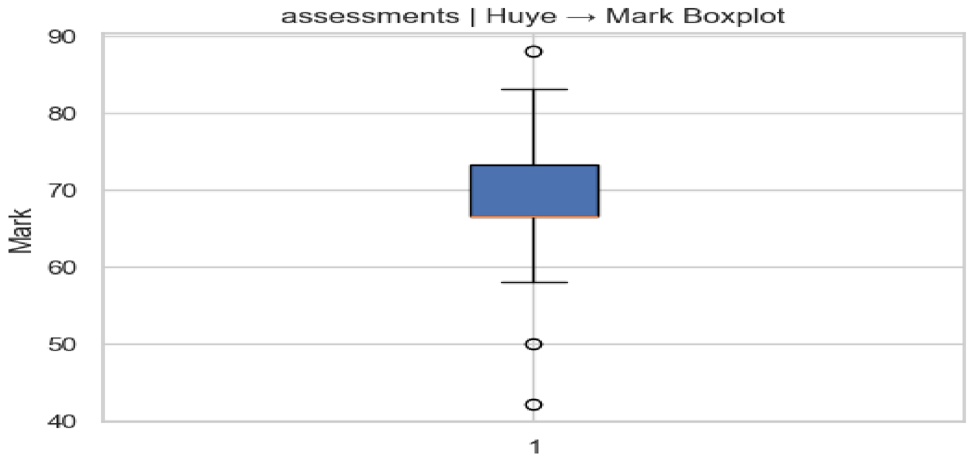


Figure 4: Outlier in Mark

4. CLEANING JUSTIFICATION

Data cleaning addressed missing values through imputation, duplicates through record selection, outliers through capping or replacement, and inconsistencies through formatting and text normalization. These steps ensured reliable integration and analysis readiness.

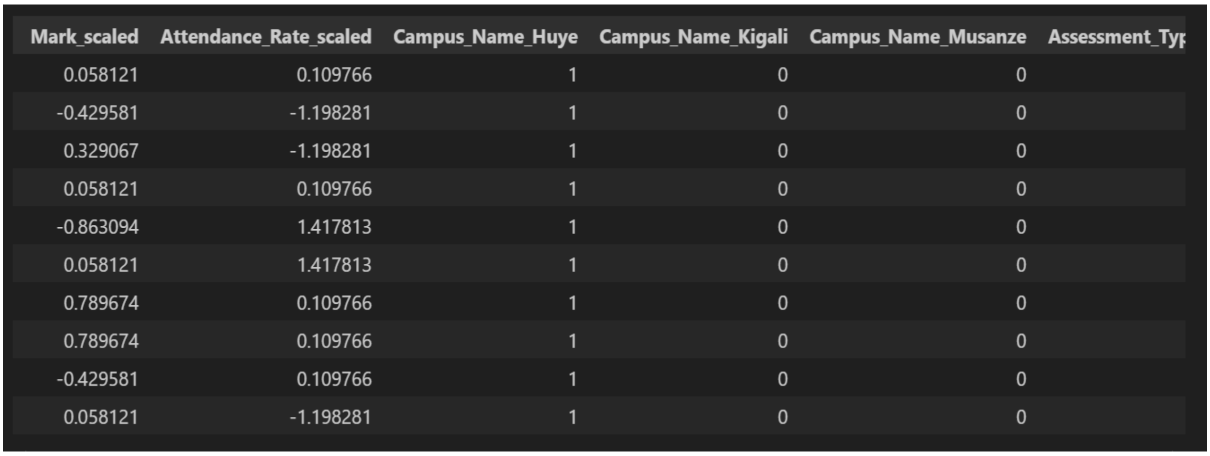
Issue	**Method Used**	**Justification**
Missing numeric values (Mark, Attendance)	Median /mean imputation	Median protects against outliers, mean used for symmetric distribution like Attendance.
Missing categorical values	Mode imputation per campus	Maintains realistic categorical distribution
Date missing	Mode imputation	Ensures consistent dates

Duplicates	Keeps latest record	Latest record assumed most accurate; resolves repeated Student_IDs
Outliers (Mark)	Capping or invalid → NaN → imputed	Preserves data distribution while removing unrealistic extremes
Text/Format inconsistencies	Strip spaces, uppercase IDs and codes, title () for names	Standardizes merge keys and prevents mismatches
Noisy characters	Regex removal of non-alphanumeric characters	Ensures clean textual data for merging and analysis

Table 2: Data cleaning methods

5. TRANSFORMATION

During transformation, numeric features such as marks and attendance were standardized using Z-score scaling. Categorical variables were encoded for analysis, and marks were grouped into performance bands (Fail, Pass, Credit, Distinction). These steps produced a consistent silver dataset ready for integration.



Mark_scaled	Attendance_Rate_scaled	Campus_Name_Huye	Campus_Name_Kigali	Campus_Name_Musanze	Assessment_Typ
0.058121	0.109766	1	0	0	
-0.429581	-1.198281	1	0	0	
0.329067	-1.198281	1	0	0	
0.058121	0.109766	1	0	0	
-0.863094	1.417813	1	0	0	
0.058121	1.417813	1	0	0	
0.789674	0.109766	1	0	0	
0.789674	0.109766	1	0	0	
-0.429581	0.109766	1	0	0	
0.058121	-1.198281	1	0	0	

Figure 5: Sample of silver dataset

6. INTEGRATION KEYS & CONFLICT HANDLING

During the integration phase, datasets from all campuses were merged using consistent keys to ensure accurate joins. Student_ID, Campus_ID, and Course_Code were standardized by removing whitespace, converting to uppercase, and normalizing course codes. Conflicts arising from the same Student_ID across multiple campuses, such as differing names or programs, were resolved using the mode to retain the most frequent or official entry. Duplicate rows across merged datasets were removed, keeping the latest records when multiple entries existed. This ensured a clean, unified gold-level dataset without mismatched or conflicting records.

7. FEATURE ENGINEERING LIST

Additional features were engineered to support analytics, including date-derived variables (month, weekday), student performance aggregates (average mark, course count, credit totals), and risk indicators such as low attendance or repeated failures.

8. CONCLUSION

The preprocessing pipeline successfully transformed multi-campus datasets into a clean, standardized, and analysis-ready gold dataset. Missing values were imputed, duplicates and outliers addressed, and all features were standardized and encoded. Integration keys were harmonized, conflicts resolved, and meaningful features were engineered to enhance analytics. This unified dataset provides a solid foundation for accurate reporting, visualization, and predictive analysis across campuses.