RWANDA POLYTECHNIC

HUYE COLLEGE

DEPARTMENT: ICT

OPTION: IT

MODULE NAME: Data Mining and Warehousing

MODULE CODE: ITLDM801

LEVEL: 8 YEAR 4

CLASS: IT B-Tech

Rwanda Polytechnic Multi-Campus Data Quality Report

GROUP 2
Group members:

NAMES	**REG-NO**	**MARKS /100**
KALISA Augustin	25RP21655	
IRUMVA Emerance	25RP19623	
MUGABO Patrick	25RP20010	

Date: 16th Feb 2026

1. DATA QUALITY ISSUES IDENTIFIED

The initial data profiling phase revealed significant quality issues across three campus datasets (Huye, Kigali, and Musanze), encompassing student records, course enrollments, and assessment data. A comprehensive analysis identified five primary categories of data quality problems requiring immediate attention.

1.1 Missing Values

Field	Missing Count	Impact
Gender	1,160	High
Date of Birth	224	Medium
Phone Number	324	Medium
Course Code	3,777	Critical
Assessment Marks	3,610	Critical
Attendance Rate	7,683	High

1.2 Duplicate Records

- Students: 135 duplicates (3%) with same Student ID but different secondary details (e.g., phone numbers), affecting data integrity.
- Assessments: 16,273 duplicates (21.6%) for the same student-course-assessment, often with conflicting marks.

1.3 Data Outliers

- 1,509 assessment records (2%) had invalid marks outside 0–100.
- Examples: -5, -10, 120, 150, 999.
- Likely caused by data entry errors and affected analysis accuracy.

1.4 Inconsistent Formatting

Field	Inconsistencies
Gender	M, F, Male, Female, m, f (mixed case and formats)
Level	L4, L5, L6, Level 4, Level 5, Level 6
Semester	Semester 1, Semester 2, SEM1, SEM2, S1, S2
Course Code	CS101, CS-101, cs101 (inconsistent delimiters and case)

1.5 Noisy Data

- Student names had extra spaces, inconsistent capitalization, and irregular formatting.
- Program names also showed random capitalization issues.
- These errors complicated matching, integration, and accurate data linkage.

. 2. DATA CLEANING METHODOLOGY AND JUSTIFICATION

2.1 Missing Value Treatment

- Gender inferred using naming conventions (missing reduced to 0).
- Records with missing Course Codes removed.
- Missing marks excluded to avoid bias.
- Phone numbers kept as null (non-critical field).
- Attendance imputed using campus/year median.

2.2 Duplicate Resolution

- Students: kept most complete and recent record (135 fixed).
- Assessments: kept highest mark per student-course (16,273 duplicates removed).

2.3 Outlier Correction

- Removed 1,509 invalid marks (<0 or >100).
- Ensured all scores fall within 0–100.

2.4 Format Standardization

- Standardized Gender, Level, Semester, and Course Codes.
- Unified naming formats to prevent mismatch errors.

2.5 Text Field Cleaning

- Removed extra spaces and fixed capitalization.
- Ensured accurate matching during data integration.

2.6 Cleaning Impact Summary

Metric	**Before Cleaning**	**After Cleaning**
Student Records	4,635	4,500
Course Records	24,652	23,719
Assessment Records	75,401	56,896
Total Records Removed	—	19,573
Data Quality Score	72%	100%

3. DATA INTEGRATION ARCHITECTURE

The integration phase consolidated cleaned datasets from three campuses into a unified analytical database. A carefully designed key structure and conflict resolution framework enabled seamless cross-campus data linkage while preserving referential integrity.

3.1 Primary Integration Keys

The integration strategy employed a hierarchical key structure leveraging natural business identifiers:

Join Type	Key Fields	Purpose
Students -Courses	Student_ID	Link student demographics to enrolled courses
Courses - Assessments	Student_ID, Course_Code	Connect course enrollments to assessment results
Full Integration	Student_ID, Course_Code, Academic_Year, Semester	Create complete student performance records

3.2 Conflict Resolution Framework

During data integration, conflicts were resolved using a structured framework. Name spelling variations were handled through fuzzy matching (85% similarity threshold), with the most recent record selected as authoritative and about 2% flagged for manual review. Course code mismatches were mostly fixed through normalization, while remaining issues were corrected using a master course catalog, excluding only 0.3% invalid or outdated codes. Duplicate columns from dataset joins were resolved by keeping the courses table as the authoritative source for shared fields, applying suffix labels during merging, and removing redundant columns after integration.

3.3 Integration Architecture

The integration followed three stages: first, a LEFT JOIN between Students and Courses preserved all students, including those not enrolled; second, a LEFT JOIN with Assessments maintained enrollment records even without assessment data; third, duplicate columns were consolidated, names standardized, and referential integrity validated. The final gold dataset contained 56,896 records covering 4,500 students across 10 courses; with complete academic history and 100% key match accuracy, ensuring no integrity issues. Level granularity. Postintegration validation confirmed 100% key match success rate and zero referential integrity violations.

4. FEATURE ENGINEERING AND ENRICHMENT

Following integration, a comprehensive feature-engineering pipeline transformed raw data into analytically rich representations. The engineered features span demographic encoding, temporal extraction, academic performance indicators, and behavioral flags designed to support predictive modeling and descriptive analytics.

4.1 Categorical Encoding Features

One-hot encoding was applied to high-cardinality categorical variables to enable machine learning model compatibility:

Feature Category	Generated Features
Campus	Campus_Rwanda Polytechnic Huye, Campus_Rwanda Polytechnic Kigali, Campus_Rwanda Polytechnic Musanze
Program	Program_Civil Engineering, Program_Computer Science, Program_Electronics, Program_Information Technology, Program_Mechanical Engineering, Program_Software Engineering
Assessment Type	Assessment_Assignment, Assessment_Cat, Assessment_Final Exam, Assessment_Final Project, Assessment_Quiz

4.2 Temporal Features

Temporal features were created by breaking dates into detailed components to analyze patterns. These include **assessment_month (1–12)** for seasonality, **assessment_weekday (0–6)** for day-based trends, and an **is_weekend_assessment** flag to compare weekend and weekday performance..

4.3 Performance Indicators

Student-level aggregate features provide comprehensive performance summaries:

Feature	Description
student_avg_mark	Mean mark across all assessments for each student
student_max_mark	Highest mark achieved by student across all assessments
student_fail_count	Count of failed assessments (mark < 50) per student
student_course_count	Total number of unique courses enrolled by student
student_total_credits_earned	Cumulative credits earned from passed courses
credits_earned	Credits earned for individual assessment (binary: earned or not)

4.4 Behavioral Flags and Risk Indicators

Binary flags enable cohort identification and intervention targeting:

Flag	Condition	Prevalence
is_fail	Assessment mark below 50	28%
low_attendance_flag	Attendance rate below 75%	22%
is_at_risk	Average mark < 60 AND low attendance	18%
high_performer	Average mark ≥ 80	15%
struggling_student	Fail count ≥ 3	12%

4.5 Derived Categorical Features

Derived features include Performance_Band (Credit, Pass, Fail) with numeric encoding, Attendance_Category (Excellent, Good, Poor) to measure engagement, and a standardized Attendance_Rate_Scaled (z-score) to support scale-sensitive machine learning models.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Key Achievements

The project built a complete data integration pipeline for Rwanda Polytechnic’s three campuses, transforming 104,688 raw records into 56,896 high-quality records and improving data quality from 72% to 96%; it removed 19,573 invalid records, ensured 100% referential integrity, and created 15-engineered features to support advanced analytics and predictive modeling.

6.2 Final Dataset Characteristics

Metric	**Value**
Total Records	56,896
Unique Students	4,500
Unique Courses	10
Total Features	38
Missing Values	< 1% across all fields
Data Quality Score	100%

6.3 Key Deliverables

- Bronze Layer: Raw combined datasets from three campuses with metadata tracking
- Silver Layer: Cleaned and standardized datasets ready for integration
- Gold Layer (Integrated): Unified dataset combining students, courses, and assessments
- Gold Layer (Features): Enriched dataset with 45+ engineered features for analytics
- Gold Layer (Reduced): Optimized dataset with redundant features removed

6.3 Recommendations

- Automated Data Validation: Implement real-time validation rules at data entry points
- Master Data Management: Establish centralized student and course master records

- Continuous Monitoring: Deploy data quality dashboards tracking completeness, accuracy, and timeliness
- Predictive Analytics Deployment: Build and deploy student success prediction models
- Data Governance Framework: Establish policies, standards, and procedures for ongoing data quality

6.5 Closing Statement

This project demonstrates that systematic data preprocessing and integration methodologies can transform fragmented, low-quality multi-source data into a unified, high-quality analytical asset. The resulting dataset positions Rwanda Polytechnic to make data-driven decisions supporting student success, operational efficiency, and institutional excellence. The methodologies and frameworks established are replicable and scalable, providing a template for similar multi-campus integration initiatives across educational institutions.