DATA QUALITY REPORT

1. Introduction

The purpose of this data quality audit is to evaluate and improve the reliability, consistency, and readiness of the data used across the student recruitment and admissions workflow. The organization relies on three primary datasets that capture interactions with prospective students from their initial outreach stage through their admission application process and campaign grouping:

1. **Outreach Dataset:**
   Contains records of call interactions conducted by the outreach engagement team to inform, support, and convert prospective students.

2. **Application Dataset:**
   Contains personal, academic, and contact information submitted by applicants during the admissions process.

3. **Campaign Dataset:**
   Provides contextual information describing the campaigns and grouping logic applied to outreach lists.

These datasets originated from different systems and were captured at different stages of the student engagement funnel. As a result, the data contained inconsistencies, missing values, formatting variations, and key alignment issues that affected their ability to be integrated and analyzed directly.

The aim of this audit was to:

- Identify structural and content-level data quality issues.
- Perform systematic cleaning to improve accuracy and usability.
- Standardize key identifiers used to link records across datasets.
- Merge the cleaned datasets into one coherent, analysis-ready dataset.

The outcomes of this audit ensure that the final merged dataset can be confidently used for reporting, analysis, campaign effectiveness evaluation, and strategic decision-making.

2. Dataset-Level Audit

A. Outreach Dataset

Description:
Contains call interaction logs for prospective applicants contacted by outreach team members.

Key Issues Identified

Issue Type	Observations	Impact
Missing Values	Some records lacked call date/time or caller information.	Reduced accuracy in call timeline analysis.
Inconsistent Name Formatting	Names contained extra spaces, mixed cases, or special characters.	Reduced matching consistency across datasets.
Duplicate Outreach Records	Some records were exact duplicates due to system export repetition.	Risk of double-counting interactions.

Cleaning Actions Applied

Action	Description	Result
Standardized Text Fields	Trimmed spaces and converted Name/University to title case.	Improved readability and matching accuracy.
Converted Dates	Converted Received_At to standard datetime format.	Enabled accurate chronological sorting.
Removed Exact Duplicates	Dropped only rows that were completely identical.	Preserved legitimate multiple outreach attempts.
Ensured Reference_ID Integrity	Verified all Reference_ID values were correctly formatted and not artificially generated.	Maintained reliable key for dataset merging.

B. Application Dataset

Description:
Contains applicant-submitted academic and contact details during university application.

Key Issues Identified

Issue Type	Observations	Impact
Incorrect Country Entries	Some applicants entered email addresses in the Country field.	Affected geographical segmentation and reporting.
Inconsistent Contact Format	Phone numbers had mixed formatting, symbols, or missing country code.	Limited reliability of contact number analytics.
Text Case Variation	Fields contained inconsistent casing (UPPER, lower, Title case).	Impaired match accuracy with outreach dataset.

Cleaning Actions Applied

Action	Description	Result
Corrected Country Field	Replaced email-based entries with Unknown.	Eliminated invalid geographic values.
Standardized Phone Format	Cleaned the phone field by removing special characters.	Improved contact consistency.
Normalized Text Fields	Applied .str.strip() and case normalization.	Reduced noise in text comparison.

Standardized App_ID	Converted App_ID to uppercase and removed whitespace.	Enabled correct merging with Outreach dataset.

C. Campaign Dataset

Description:
Contains metadata describing the outreach campaigns associated with each record.

Key Issues Identified

Issue Type	Observations	Impact
Inconsistent ID Format	Campaign ID (ID) contained extra spaces or inconsistent casing.	Prevented accurate merging with Outreach dataset.
Redundant Descriptive Fields	Certain text notes repeated campaign category information.	Increased dataset size without analytical benefit.

Cleaning Actions Applied

Action	Description	Result
Standardized ID Field	Trimmed and converted Campaign ID to uppercase.	Ensured proper merge alignment.
Removed or Consolidated Redundant Columns	Keep only useful classification and intake attributes.	Simplified campaign analysis.

---

3. Merged Final Dataset

After individually cleaning the Outreach, Application, and Campaign datasets, the next step was to integrate them into a single consolidated dataset that reflects the complete student interaction journey—from initial outreach call to formal application submission and campaign grouping.

The merging process was carried out in two main stages:

Merge Step	Datasets Involved	Key(s) Used for Matching	Join Type	Purpose
Step 1	Outreach + Application	Reference_ID (Outreach) ↔ App_ID (Application)	**Left Join**	To attach applicant demographic and application details to outreach call records
Step 2	(Outreach + Application) + Campaign	Campaign_ID (Outreach) ↔ ID (Campaign)	**Left Join**	To map each outreach call to the campaign structure and category

Reason for Using Left Joins

The Outreach dataset is the primary source capturing all students contacted.
Not all contacted students submit applications; therefore:

- Using **Left Join** ensures **no outreach interaction is lost**.
- Application and Campaign details appear **only where matching records exist**.

This preserves the true engagement funnel:
Contacted → Interested → Applied → Admitted

Final Validation & Quality Checks Performed

Check Performed	Description	Result
Match Rate Review	Verified how many Reference_ID values matched to App_ID	Natural partial match observed (Expected: Not all outreach leads apply)
No Artificial ID Creation	Confirmed no new or synthetic IDs were created during processing	All IDs remained original and traceable
Retention of Multiple Outreach Attempts	Ensured multiple call records for the same student were not removed	All engagement attempts preserved for call performance analysis
Missing Values Handling	Fields where matching data was not available were filled with neutral placeholders like Unknown (instead of blank or incorrect assumptions)	Prevents accidental misinterpretation during reporting
Campaign Label Consistency Check	Ensured Campaign names and categories remained correctly mapped after merge	Campaign data aligned and complete where applicable

---

Outcome of the Merge

The final merged dataset:

- Represents a **complete 360° engagement view** of each student.

- Maintains the **accuracy of call history**, application progress, and campaign allocation.

- Is now **standardized, analysis-ready, and traceable** back to original data sources.

4. Summary & Data Readiness

The three datasets have been successfully:

- Cleaned for formatting consistency

- Standardized on join keys

- Merged into a single structured dataset

The final dataset is now:

- **Accurate** – No artificial data was introduced.

- **Consistent** – Standard formats across key fields.

- **Analysis-Ready** – Suitable for reporting, dashboarding, and modeling.

Some non-critical missing values remain (e.g., unknown country), which should be **considered during interpretation**, but they do **not limit the use of the dataset**.

Overall Readiness: The dataset is ready for analysis and insights generation.