# UNSTRUCTURED TO STRUCTURED DATA

**1. Overview**

The raw dataset was provided in multiple Excel files containing several sheets with inconsistent formats, column names, and missing values. The goal was to convert this unstructured data into a single, clean, and structured dataset suitable for analysis and visualization.

This transformation was performed using **Microsoft Excel – Power Query**.

**2. Challenges in Raw Data**

The raw data had the following issues:

- Data distributed across multiple Excel sheets
- Different column names representing the same information (e.g., Email, Email Address)
- Multiple email and mobile number columns created after appending
- Presence of null and blank rows
- Duplicate student records
- Inconsistent data types across columns

**3. Step-by-Step Data Structuring Process**

**Step 1: Importing Raw Data**

- Loaded all Excel sheets using Get Data → Excel Workbook
- Selected each relevant sheet and opened it in Power Query Editor

**Step 2: Standardizing Individual Sheets**

For each sheet:

- Promoted the first row as headers
- Renamed columns to a standard format (Name, Email, Mobile, Age, Branch, Course, Institution, Enrollment Number)
- Converted column data types (Text, Whole Number, etc.)
- Removed completely empty rows

**Step 3: Combining Multiple Sheets**

- Used Append Queries to vertically merge all sheets into one consolidated table
- Ensured that column names matched across all sheets before appending
- This created a single dataset containing all student records

**Step 4: Resolving Duplicate Columns**

After appending, multiple columns such as Email and Mobile were created due to inconsistent naming.
To resolve this:

- Created new custom columns using conditional logic to select non-null values
- Example logic:
  **If primary column is null, take value from secondary column**
- Removed the original duplicate columns after verification

**Step 5: Handling Missing Values**

- Filtered out rows where critical fields (Name, Email) were null
- Ensured no incomplete student records remained

**Step 6: Removing Duplicate Records**

- Used Remove Duplicates on the Email column
- This ensured each student record appeared only once

**Step 7: Final Validation**

- Verified total row count after cleaning
- Checked for remaining nulls in key fields
- Confirmed consistent data types across all columns

### 4. Final Structured Dataset

The final output is a clean, structured table with:

- One row per student
- Standardized columns
- No duplicate or incomplete records
- Ready for visualization and further analysis

### 5. Outcome

The transformation process successfully converted raw, unstructured Excel data into a structured dataset that supports accurate reporting and visualization in Power BI.