

## 1. Data Ingestion and Setup

I began by importing the required Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn. I then loaded all the datasets — Raw Data 1, Raw Data 2, and the provided template file.

As an initial step, I inspected the datasets by checking their column names, shapes, and basic structure to understand the format, completeness, and consistency of the data.

## 2. Course and Branch Standardisation

For Raw Data 1, the Course and Branch columns contained missing values. Based on the document name (“INFORMATION TECH”), I inferred that these students belong to the Information Technology stream.

Therefore, I filled all missing Course and Branch values with “**Information Technology**” following the template instruction to use file or tab names as hints where data is missing.

## 3. Dataset Consolidation

After preparing both datasets, I merged Raw Data 1 and Raw Data 2 into a single combined dataset using concatenation. This created a unified dataset for further processing and cleaning.

## 4. Phone Number Cleaning and Validation

Phone number was treated as the primary identifier for a job seeker. I cleaned the phone number column by fixing scientific notation (float values), converting all values to integers, and validating that each phone number contained exactly 10 digits.

Rows with invalid or missing phone numbers were removed to ensure that every record represents a contactable job seeker profile.

## 5. Age Cleaning and Validation

Age values were first converted from float format to integers.

I then applied a working-age filter and retained only records with age between **16 and 60 years**, removing invalid entries such as ages 2, 3, or 15 which are not valid for employment.

## 6. Email Validation

Email IDs were cleaned and validated by checking whether they contained the “@” symbol.

Invalid or missing emails were marked as “**Not Provided**” since email is an optional field but still useful for contact and verification.

## 7. Text Formatting and Normalisation

Text formatting was standardised for the following columns:

- Name
- Email
- Permanent Address
- Address of Associated Institute (with map link)
- Course
- Branch
- Institution
- Notes

Leading and trailing spaces were removed and formatting inconsistencies were fixed.

The Notes column was found to be empty across all records and was retained as an optional field.

## 8. Template Mapping and Column Reordering

The dataset was then aligned with the provided template.

Columns were reordered exactly as per the template structure to make the file compatible for bulk upload into the ONEST Blue Dot system.

A **“Leave blank”** column was added as required by the template.

## 9. Institute Standardisation

For rows where the Institution field was missing, I filled the value as **“Govt Polytechnic Ghaziabad”** based on the instruction provided in the template and the source of the dataset.

## 10. Duplicate Handling

Duplicate job seeker records were identified using phone number as the unique identifier.

In cases where multiple registrations existed for the same phone number, only one profile was retained to ensure a single Blue Dot per job seeker.

## 11. Export and Location Linking

The final processed dataset was exported as an Excel file.

Using Excel, Google Maps links were manually added for each institute in the **“Address of Associated Institute (with map link)”** column to enable location discovery as required by the template.