

**AI Data Powered Analysis Early Internship - Sub-Group 32**

**Week-1**

**Data Dictionary Document**

**Team Members:**

| **Name** | **Role** | **Email** |
| --- | --- | --- |
| Sumaiya Tasnim | Team Lead | [sumaiyaa.tasnim.18@gmail.com](mailto:sumaiyaa.tasnim.18@gmail.com) |
| Wilgens Almonor | Project Scribe | [Wilgensalmonor@gmail.com](mailto:Wilgensalmonor@gmail.com) |
| Subashree J | Project Manager | [subashreej03@gmail.com](mailto:subashreej03@gmail.com) |
| Ayan Banik | Project Lead | [ayanbanik001@gmail.com](mailto:ayanbanik001@gmail.com) |

This document represents the Data Dictionary for the cleaned OutreachData.csv, CampaignData.csv, and ApplicantData.csv datasets. It defines each column, specifies corrected data types, and explains the purpose of every field. The document also highlights key cleaning actions taken to ensure accuracy, consistency, and readiness for analysis in the next phase.

**Data Dictionary**

1. **OutreachData.csv**

| **Column Name** | **Corrected**  **Data Type** | **Description of the Field** | **Notes on Any Changes from Original** |
| --- | --- | --- | --- |
| Reference\_ID | String | Unique identifier for each outreach record | Original contained some invalid entries (e.g., ',,,,', '/////////', '0') → Removed all non-numeric or zero Reference\_ID rows to ensure only valid IDs remain and maintain compatibility. |
| Received\_At | datetime64[ns] | Date and time when the outreach was received | Original was stored as object (string) → Converted to datetime type to allow proper date-time analysis and sorting. |
| University | String | Name of the university contacted | Original was object with potential extra spaces → Cleaned by stripping leading/trailing spaces for consistency; no value changes. |
| Caller\_Name | String | Name of the caller who made the outreach | Original contained extra spaces → Cleaned by stripping spaces to ensure uniformity in names. |
| Outcome\_1 | String | Result or status of the outreach contact | Original contained extra spaces and inconsistent formatting → Cleaned by stripping spaces; long text entries were kept as-is for detailed reporting. No modifications were made to the meaning or wording of long text entries. |
| Remark | String | Additional remarks or notes from the outreach | Original had 33,804 missing values → Cleaned by filling missing values with 'No Remark' and stripping spaces to ensure completeness and consistency. |
| Campaign\_ID | String | Identifier for the outreach campaign | Original contained extra spaces → Cleaned by stripping spaces to maintain uniform campaign codes. |
| Escalation\_Required | String | Indicates if escalation was required | Original contained inconsistent values like 'Yes, No' → Cleaned by standardizing all such cases to 'Yes' and stripping spaces to maintain consistency for reporting and analysis. |

1. **CampaignData.csv**

| **Column Name** | **Corrected Data Type** | **Description of the Field** | **Notes on Any Changes from Original** |
| --- | --- | --- | --- |
| ID | String | Unique identifier for each campaign record | Original values were all valid and unique → No changes required. |
| Name | String | Name of the campaign | Original values had some variations in naming style → No changes made; names kept as-is for reporting. |
| Category | String | Indicates whether campaign is Pre Admission or Post Admission | Original values were consistent → No changes required. |
| Intake | String | Academic year for the campaign (e.g., AY2024) | Original values were consistent → No changes required. |
| University | String | University associated with the campaign | Original values were consistent → No changes required. |
| Status | String | Current status of the campaign | Original values were consistent → No changes required. |
| Start\_Date | datetime64[ns] | Date and time when the campaign started | Original stored as object (string) → Converted to datetime format (%Y-%m-%d %H:%M:%S) to allow proper date-time analysis and sorting. |

1. **ApplicantData.csv**

| **Column Name** | **Corrected Data Type** | **Description of the Field** | **Notes on Any Changes from Original** |
| --- | --- | --- | --- |
| App\_ID | String | Unique identifier for each applicant | Original contained missing, invalid, or non-numeric entries (e.g., ',,,,', '/////////', 'na', 'naq') → Removed all invalid rows and kept only valid numeric IDs. Converted to numeric temporarily for validation, then back to object for consistency. |
| Phone\_Number | String | Applicant’s contact phone number | Original contained spaces, dashes, parentheses, special characters, extremely long invalid numbers → Cleaned to retain only digits (keeping leading '+'), rows with invalid or too long numbers replaced with 'Unknown' for compatibility. |
| Country | String | Country of the applicant | Original contained typos, lowercase, multiple countries, email-like entries, or long text → Standardized capitalization, corrected common typos, replaced invalid / multiple country entries and email-like entries with 'Unknown'. |
| University | String | Name of the applicant’s university | Original had only one unique value (Illinois Institute of Technology) → Stripped extra spaces for consistency; no value changes. |