

**Data Visualization Trainee Early Internship - Sub-Group 32**

**Week-1 Final Report**

**Report Title: “Week-1: Data Quality Report”**

**Associate Members:**

| **Name** | **Email** |
| --- | --- |
| Kolawole Oparinde | [kolawole@vempower.org](mailto:kolawole@vempower.org) |
| Khusi Capoor | [khusicapoor@vempower.org](mailto:khusicapoor@vempower.org) |
| Sudharma | [sudharma@vempower.org](mailto:sudharma@vempower.org) |
| Shruti Mishra | [shrutimishra@vempower.org](mailto:shrutimishra@vempower.org) |

**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) |

**Introduction:**

As part of the Week-1 Data Visualization Trainee Early Internship, this project focuses on auditing, cleaning, and documenting real-world outreach, campaign and applicant datasets. The goal is to transform raw data into structured, reliable, and analysis-ready datasets. By identifying inconsistencies, missing values, duplicates, and invalid entries, we ensure accuracy and clarity, enabling meaningful insights and recommendations. This deliverable demonstrates the transition from raw data exploration to preparing a robust dataset suitable for analysis and visualization.

**Objectives:**

* Conducting a systematic audit of the outreach, campaign, and application datasets.
* Identifying key data quality issues such as missing values, duplicates, inconsistent formats, and invalid entries.
* Applying appropriate data cleaning techniques to prepare the datasets for analysis.
* Documenting the cleaned datasets with a comprehensive data dictionary.
* Ensuring the datasets are ready for downstream analysis, visualization, and reporting.

**Assigned Datasets:**

| **Assigned Dataset (Raw) *Before*** | **Cleaned Dataset *After*** |
| --- | --- |
| OutreachData.csv | Cleaned\_OutreachData.csv |
| CampaignData.csv | Cleaned\_CampaignData.csv |
| ApplicantData | Cleaned\_ApplicantData |

**Tools We Will Use:**

* **Excel** – For initial exploration, data cleaning, summary statistics, and preparing the data dictionary.
* **Python (Jupyter Notebook)** – For advanced data cleaning, validation, and documenting reproducible cleaning steps.

**Expected Outcomes:**

* A Data Quality Report documenting identified issues, cleaning actions, and remaining limitations.
* Cleaned datasets ready for analysis and reporting.
* A Data Dictionary defining each column, its type, description, and notes on changes from the raw dataset.
* Demonstrated ability to transform messy, real-world data into structured, analysis-ready datasets.

**Learning Outcomes:**

1. Learning to systematically audit and validate real-world datasets.
2. Applying effective data cleaning techniques to prepare analysis-ready data.
3. Documenting datasets clearly and comprehensively for future users.
4. Building a foundation for exploratory data analysis and interactive dashboard creation in subsequent weeks.

**Dataset: OutreachData.csv**

The **OutreachData.csv** dataset contains detailed records of outreach activities conducted by Illinois Institute of Technology. It tracks interactions with prospective students, including the outcome of each contact and any remarks provided. The dataset consists of **8 columns**: Reference\_ID, Received\_At, University, Caller\_Name, Outcome\_1, Remark, Campaign\_ID, and Escalation\_Required. In total, the dataset includes **37,881 rows** and **8 columns**, providing a comprehensive view of outreach efforts over time.

**Dataset Structure:**

| **Column Name** | **Original Datatype** | **What It Represents** |
| --- | --- | --- |
| Reference\_ID | Object (TEXT) | Unique identifier assigned to each outreach record |
| Received\_At | Object (TEXT) | Date and time when the outreach record was received |
| University | Object (TEXT) | Name of the university related to the outreach (Illinois Institute of Technology) |
| Caller\_Name | Object (TEXT) | Name of the caller or staff who made the outreach |
| Outcome\_1 | Object (TEXT) | Result or outcome of the outreach interaction |
| Remark | Object (TEXT) | Additional comments or follow-up notes related to the outreach |
| Campaign\_ID | Object (TEXT) | Identifier of the campaign under which the outreach was conducted |
| Escalation\_Required | Object (TEXT) | Indicates whether the case required escalation or not |

Total Records: 37881 rows

Total Columns: 8

**Data Cleaning Process:**

The data cleaning process for the **OutreachData.csv** dataset was performed using Python programming language in Visual Studio (Jyputer Notebook). Various Python libraries and functions were used to inspect, clean, and standardize the dataset to make it ready for further analysis and dashboard reporting in Power BI for next week.

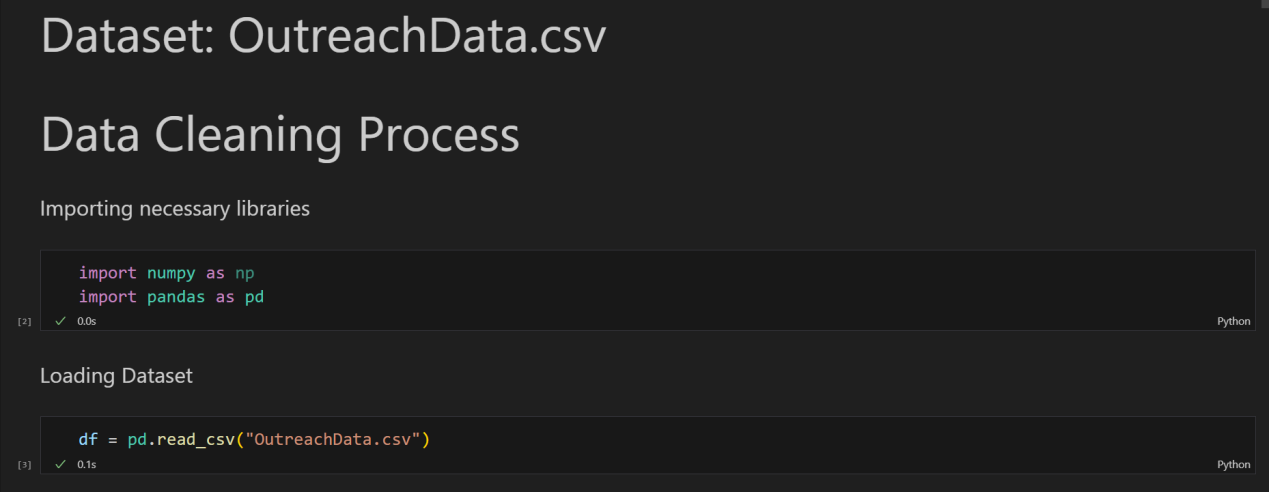
**Purpose of Data Cleaning:**

* To ensure completeness by handling missing values & detecting duplicate rows.
* To remove or correct inconsistent or invalid data.
* To standardize datatypes for seamless analysis.
* To prepare the dataset for visualizations, accurate reporting and creating dashboard in Power BI for week-2.

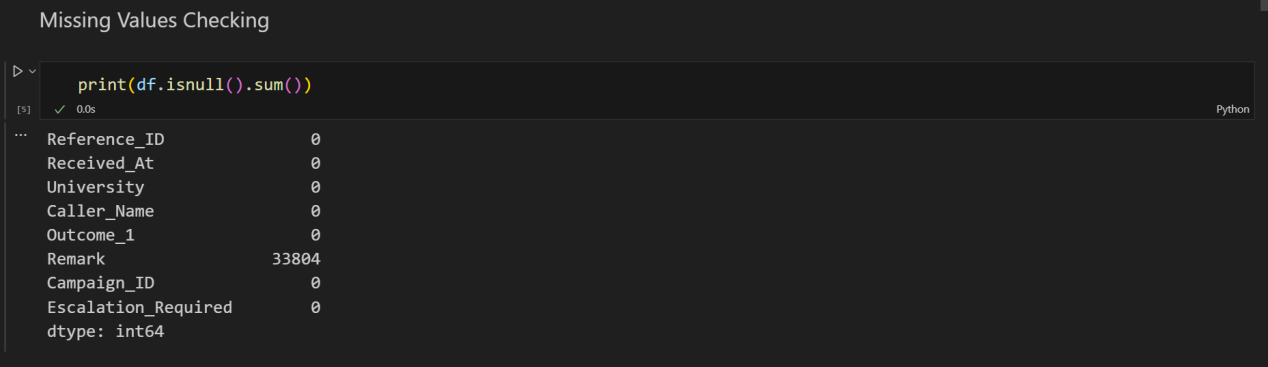
**Tools Used:**

1. Python – used for data cleaning and preprocessing. *(programming language)*
2. Pandas – for data manipulation and handling missing values. *(imported necessay library)*
3. NumPy – for numerical operations and validation. *(imported necessay library)*
4. Visual Studio Code – used as the programming environment for executing Python scripts. *(Chosen Application)*

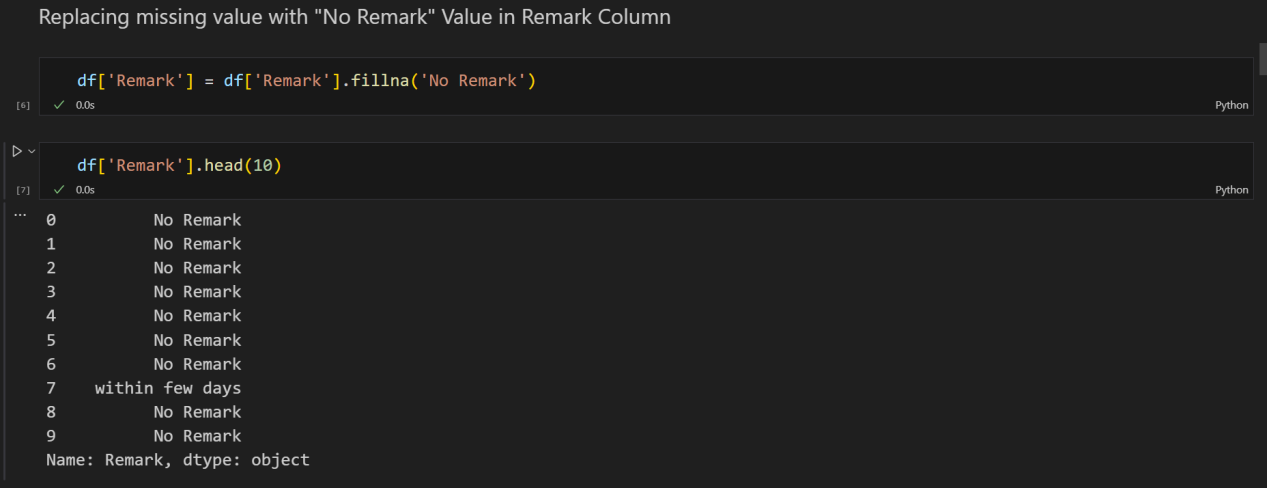
***Step-1: Importing necessary libraries & Loading Dataset***



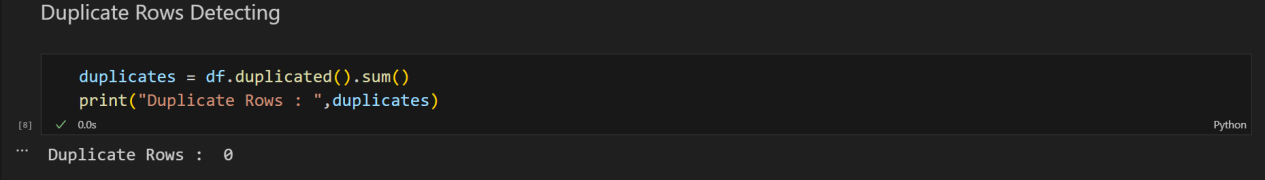
***Step-2: Checking Missing Values per Columns***



The Remark column originally contained 33,804 missing values, which were filled with 'No Remark' during the data cleaning process to ensure completeness. Then checked first 10 rows of Remark column to see if it is replaced properly.

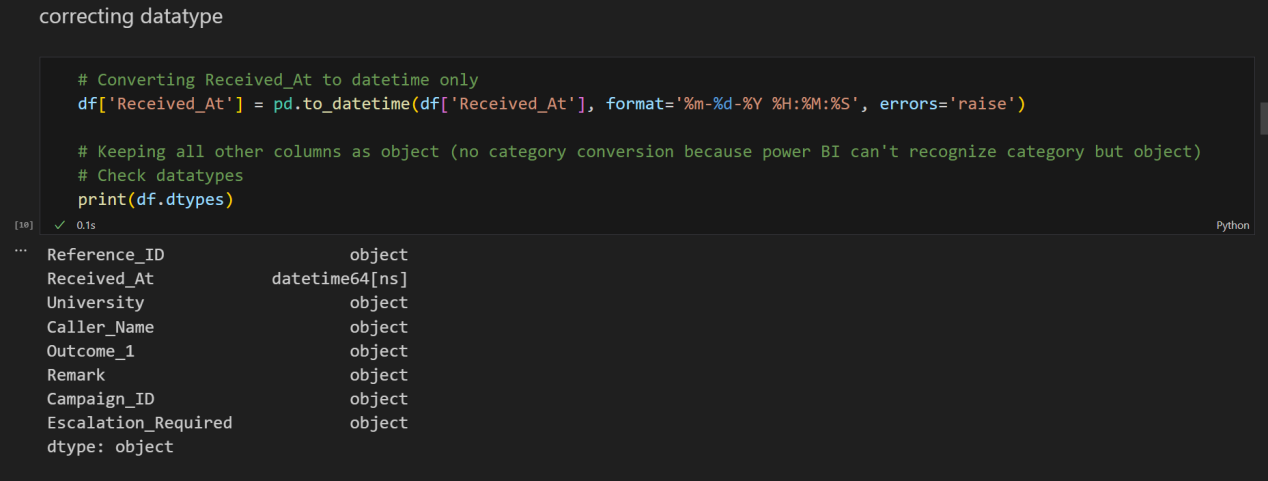


***Step-3: Detecting Duplicate Rows***



There was no duplicate rows found.

***Step-4: Correcting the datatypes with suitable dataype:***

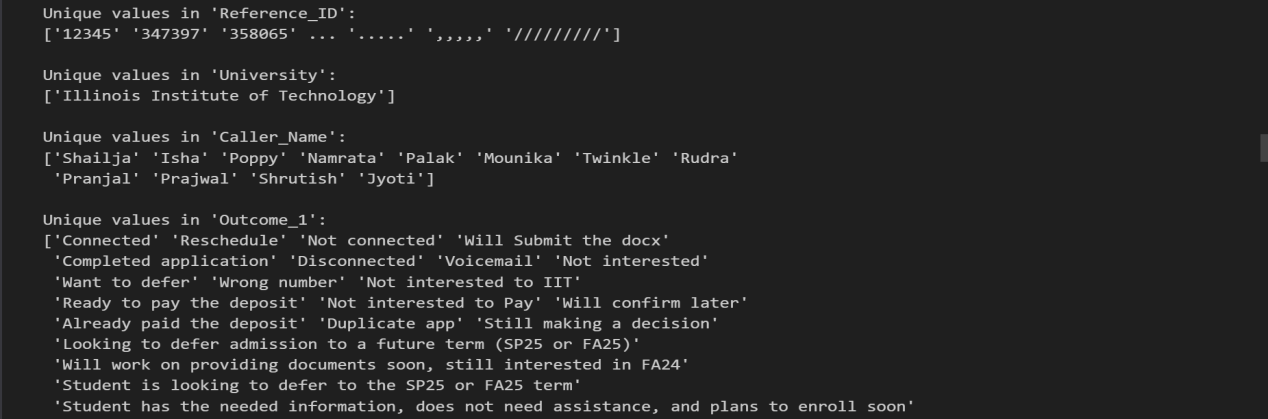


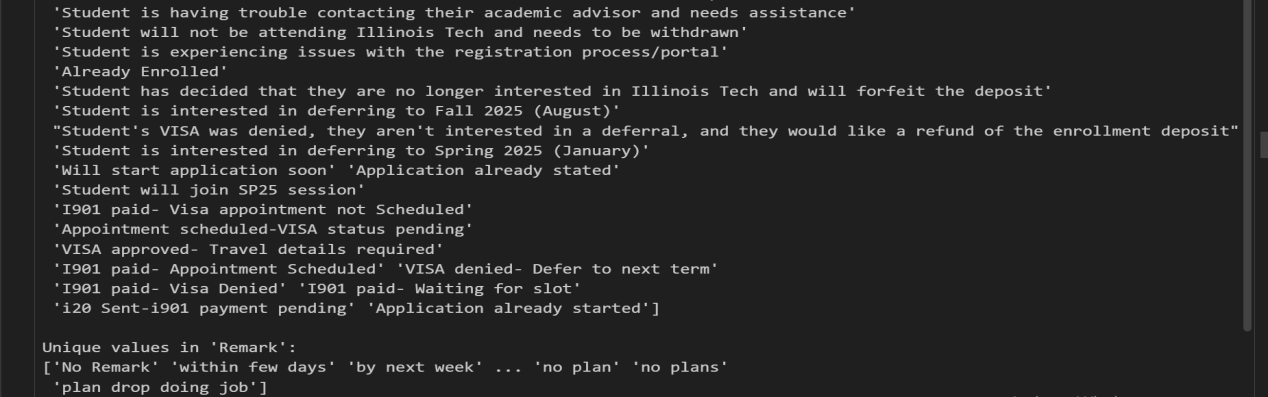
The Received\_At column was converted from object to datetime datatype to enable proper date-time analysis, while all other columns were retained as object for compatibility with Power BI.

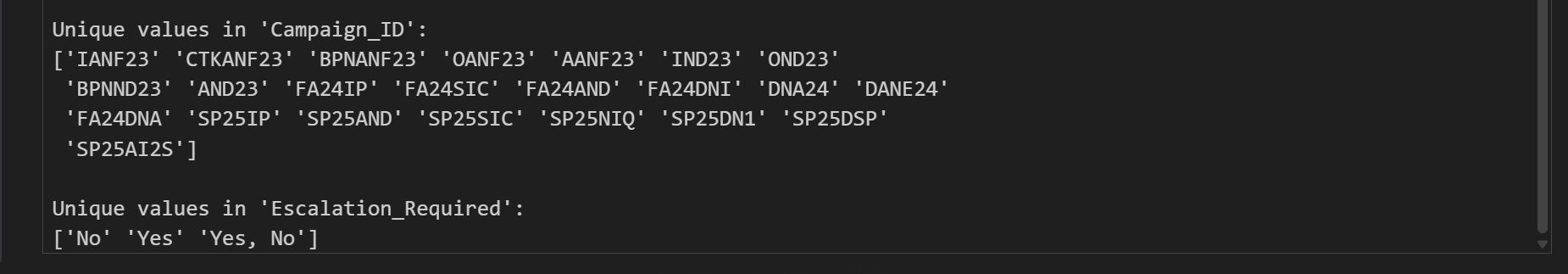
***Step-5: Checking Inconsistences in all categorical (object) columns***



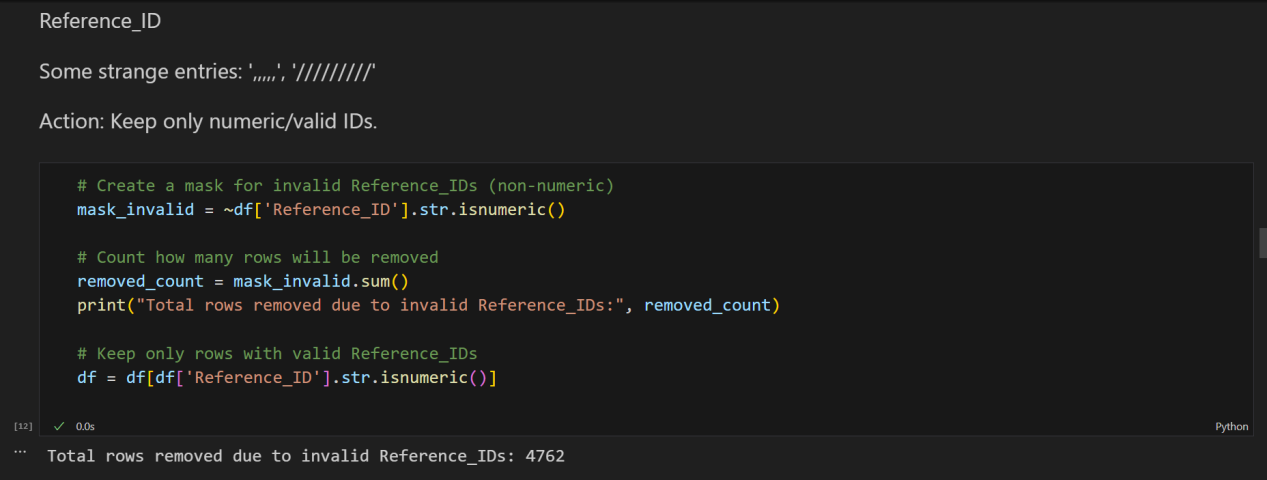
We checked the uniques values, so that we can identify if there is any invalid entries.

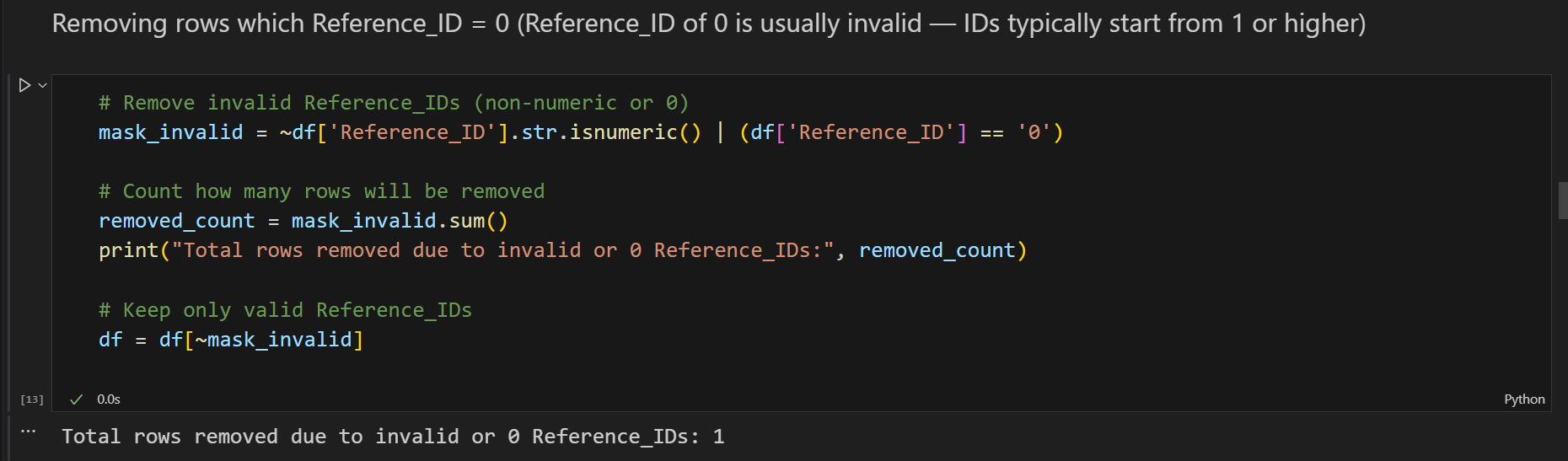






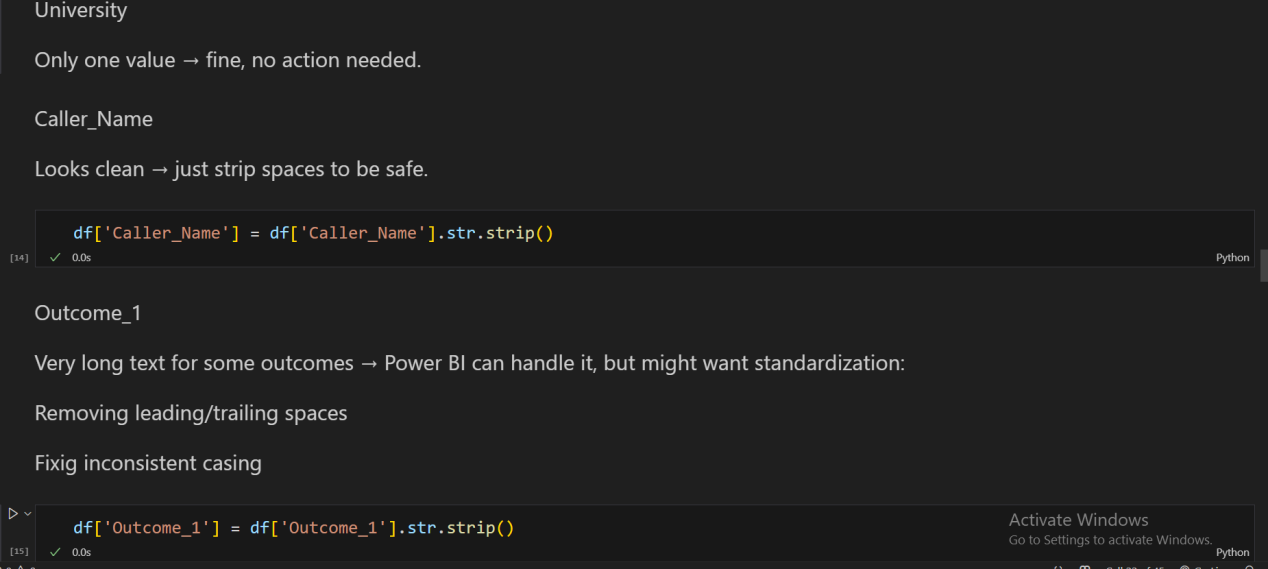
***Step-5.1: Correcting Invalid entries in Reference\_ID column***



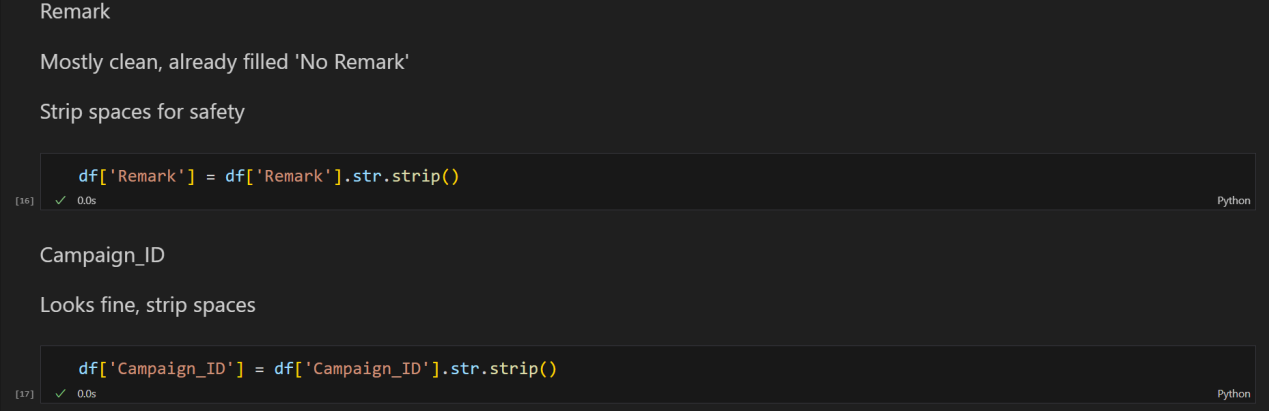


Invalid entries in the Reference\_ID column were removed to maintain consistency.

***Step-5.2: Correcting unnecessary spaces in textual values***

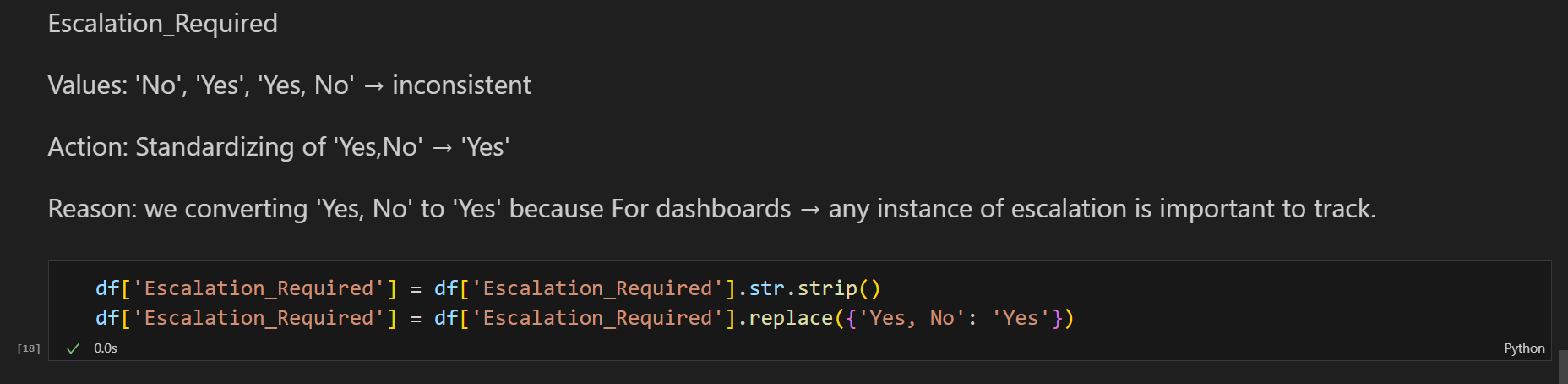


There was only one value in University column, that is “Illinois Institute of Technology”. Moreover, there was very long textual values in Outcome\_1 column, which were not invalid entries so we kept the same values. Just for safety, we ran the codes so there doesn’t remains any unnecessary spaces in the values.



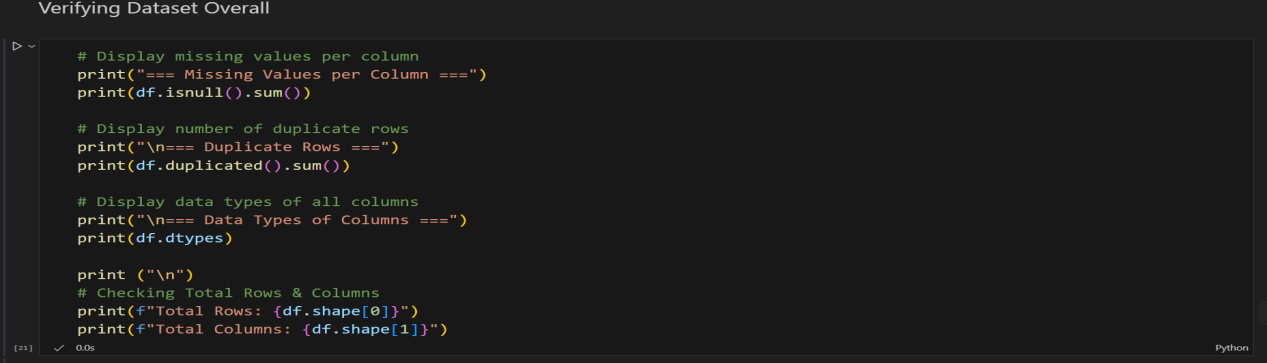
Also for column Remark & Campaign\_ID, we ran the codes so there doesn’t remains any unnecessary spaces in the values for safety. As there was no invalid entries in those columns.

***Step-5.3: Correcting Invalid entries in Escalation\_Required column & Unnecessary spaces:***



The Escalation\_Required column was standardized, replacing 'Yes, No' with 'Yes' to maintain consistency and simplify reporting & creating dashboard.

***Step-6: Verification of the Overall Dataset after data cleaning***

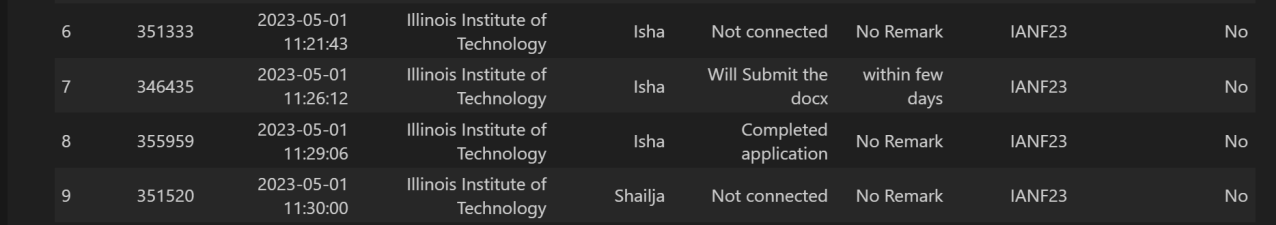
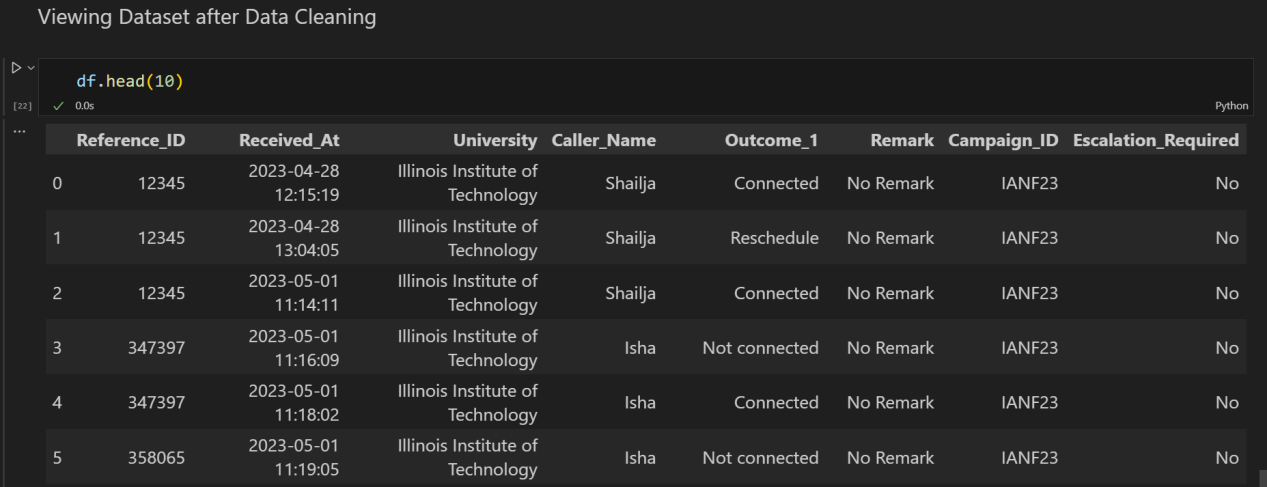




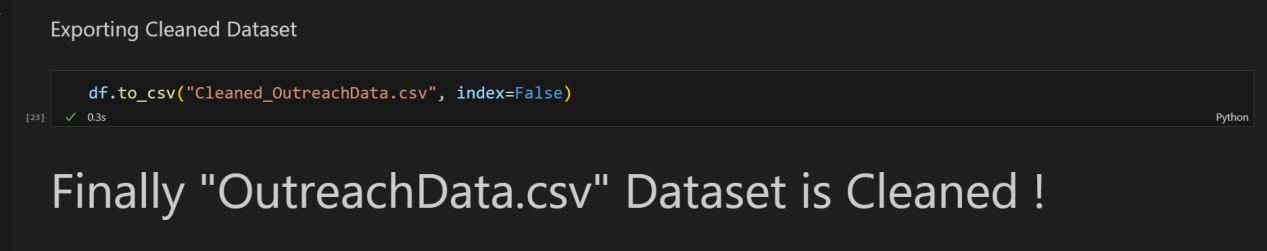
**By the output observation:**

Overall, there remains no NULL/Missing values in each column, there is no duplicate records found, all column’s dataypes are now correct, no inconsistency and total rows, columns also are fine which means data outcomes came out perfectly without changing any valid data inside it.

***Step-7: Viewing first 10 Rows of the Dataset***



***Step-8: Exported the Dataset as “Cleaned\_OutreachData.csv”***



**Data Cleaning Summary:**

| **Columns** | **Action Taken** | **Code / Methodology Used** | **Observations / Reason** |
| --- | --- | --- | --- |
| All columns | Checked total rows & columns | df.shape | Verified dataset size (37,881 rows × 8 columns). |
| All columns | Checked missing values | df.isnull().sum() | Identified missing values, especially in Remark. |
| All columns | Checked duplicate rows | df.duplicated().sum() | No duplicates found. |
| Received\_At | Converted to datetime | pd.to\_datetime(df['Received\_At']) | Ensures proper date-time analysis. |
| Remark | Filled missing values with 'No Remark' | df['Remark'].fillna('No Remark', inplace=True) | Completeness for reporting (33,804 missing values handled). |
| Object columns (Caller\_Name, Outcome\_1, University, Campaign\_ID, Escalation\_Required) | Stripped leading/trailing spaces | df[col] = df[col].str.strip() | Removes accidental whitespace. |
| Escalation\_Required | Standardized values | df['Escalation\_Required'].replace('Yes, No', 'Yes', inplace=True) | Replaced inconsistent value 'Yes, No' with 'Yes'. |
| Reference\_ID | Removed invalid Reference\_ID contained rows | mask\_invalid = ~df['Reference\_ID'].str.isnumeric() | Invalid entries were removed to ensure consistency. |
| All object columns | Checked unique values | df[col].unique() | Verified consistency and corrected minor issues. |
| All columns | Verified missing values, duplicates, datatypes | Combined checks for isnull(), duplicated(), and dtypes | Final sanity check before export. |
| The dataset | Exported as cleaned CSV file format | df.to\_csv('Cleaned\_OutreachData.csv', index=False) | “Cleaned\_OutreachData.csv” ready for visualization and dashboard use |

**Data Quality Result summary:**

| **Operation** | **Before** | **After** |
| --- | --- | --- |
| Total Rows | 37,881 | 33,118 |
| Total Columns | 8 | 8 |
| Missing Values in Remark | 33,804 | 0 |
| Duplicate Rows | 0 | 0 |
| Reference\_ID Invalid / Non-numeric | Present (e.g., ',,,,', '/////////', '0') | Removed |
| Received\_At Datatype | object | datetime64[ns] |
| Escalation\_Required Inconsistencies | 'Yes, No' present | Standardized to 'Yes' |
| Other Object Columns | Possible extra spaces | Stripped leading/trailing spaces |

**Data Dictionary:**

| **Column Name** | **Corrected**  **Data Type** | **Description of the Field** | **Notes on Any Changes from Original** |
| --- | --- | --- | --- |
| Reference\_ID | Object (TEXT) | 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 | Object (TEXT) | 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 | Object (TEXT) | Name of the caller who made the outreach | Original contained extra spaces → Cleaned by stripping spaces to ensure uniformity in names. |
| Outcome\_1 | Object (TEXT) | 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 | Object (TEXT) | 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 | Object (TEXT) | Identifier for the outreach campaign | Original contained extra spaces → Cleaned by stripping spaces to maintain uniform campaign codes. |
| Escalation\_Required | Object (TEXT) | 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. |

**Dataset Overview after cleaning:**

* No null or missing values in any column.
* No duplicate rows.
* Correct datatypes for all columns.
* No inconsistencies in the data.
* Original valid data remained unchanged.
* Invalid entries were replaced without dropping unnecessary rows.

**Dataset: CampaignData.csv**

The CampaignData is a dataset that contains detailed records of campaign activities conducted by the University of Technology. The dataset consists of 7 columns: ID, Name, Category, Intake,

University, Status, Start\_date. In total, the dataset has 23 rows and 7 columns in which it provides information about campaign activities.

**Dataset Structure:**

| Column Name | Original Datatype | Description / What it Represents |
| --- | --- | --- |
| ID | Object (TEXT) | Unique identifier for each campaign record |
| Name | Object (TEXT) | Name of the campaign record |
| Category | Object (TEXT) | Categorizes the data by admission type (post-admission or pre-admission) |
| Intake | Object (TEXT) | Admission year associated with the campaign |
| University | Object (TEXT) | Name of the university related to the campaign data (e.g., Illinois Institute of Technology) |
| Status | Object (TEXT) | Completion status of the campaign data |
| Start\_Date | Object (TEXT) | Date and time when the campaign started or was received |

Total Records: 23

Total Columns: 7

**Data Cleaning Process:**

The data cleaning process for the Campaindata.csv was performed using excel. The data was cleaned using standardizing format and finding inconsistencies. I’ve checked for duplicate values that could skew my analysis.

**Purpose of Data Cleaning:**

* Removing duplicate and missing values
* Maintaining consistency of data
* Preparing the data for analysis

**Data Quality Result summary:**

| **Operation** | **Detected / Before** | **Correction / After** |
| --- | --- | --- |
| Missing values | None (0 nulls in all columns) | No change |
| Duplicate rows | None | No change |
| ID column | 23 unique IDs, text | No change |
| Name column | 23 unique names | No change |
| Category column | 2 unique values (Post Admission, Pre Admission) | No change |
| Intake column | 1 unique value (AY2024) | No change |
| University column | 1 unique value (Illinois Institute of Technology) | No change |
| Status column | 1 unique value (Completed) | No change |
| Start\_Date column | 17 values in mixed text format (3/20/2024 0:00 etc.) | Converted to datetime64[ns] in %Y-%m-%d %H:%M:%S format |
| Saved CSV | N/A | Saved as Cleaned\_CampaignData.csv |

**Data Dictionary:**

| **Column Name** | **Corrected Data Type** | **Description of the Field** | **Notes on Any Changes from Original** |
| --- | --- | --- | --- |
| ID | Object (TEXT) | Unique identifier for each campaign record | Original values were all valid and unique → No changes required. |
| Name | Object (TEXT) | Name of the campaign | Original values had some variations in naming style → No changes made; names kept as-is for reporting. |
| Category | Object (TEXT) | Indicates whether campaign is Pre Admission or Post Admission | Original values were consistent → No changes required. |
| Intake | Object (TEXT) | Academic year for the campaign (e.g., AY2024) | Original values were consistent → No changes required. |
| University | Object (TEXT) | University associated with the campaign | Original values were consistent → No changes required. |
| Status | Object (TEXT) | 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. |

**Dataset Overview after cleaning:**

* No null or missing values in any column.
* No duplicate rows.
* Correct datatypes for all columns.
* No inconsistencies in the data.
* Original valid data remained unchanged.

**Dataset: ApplicantData.csv**

There are 4 columns in the dataset.

1. App\_ID
2. Country
3. University
4. Phone\_Number

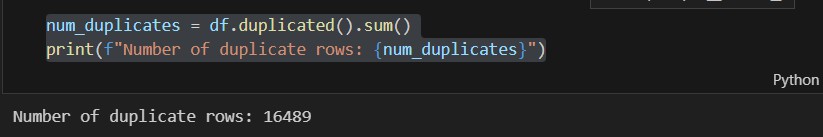
App\_ID, Country and Phone\_Number columns have inconsistent, invalid data. So, I need to clean this data to make this data usable.

We used following tools and language to clean this data:

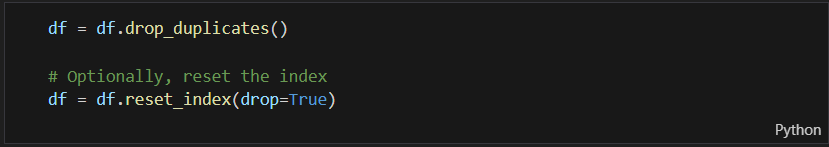
1. Python libraries pandas, os, re.
2. Kaggle notebook

Step 1: Duplicate Rows Handling

* + Count the number of duplicate rows.



* + Keep the first occurrence of the duplicate rows and remove other rows.



Step 2: App\_ID Column Cleaning

First, I went through all the data manually in the excel. I tried to see all the data. I find that App\_ID columns have mixed data. Example,

“351333”, “ ”, “Lebanon”, “2330335962”, “A20451333” “12345” etc.

The most common format of the App\_ID column data is six digits number. The other format is invalid. Also, there is another problem that some of the App\_ID data are

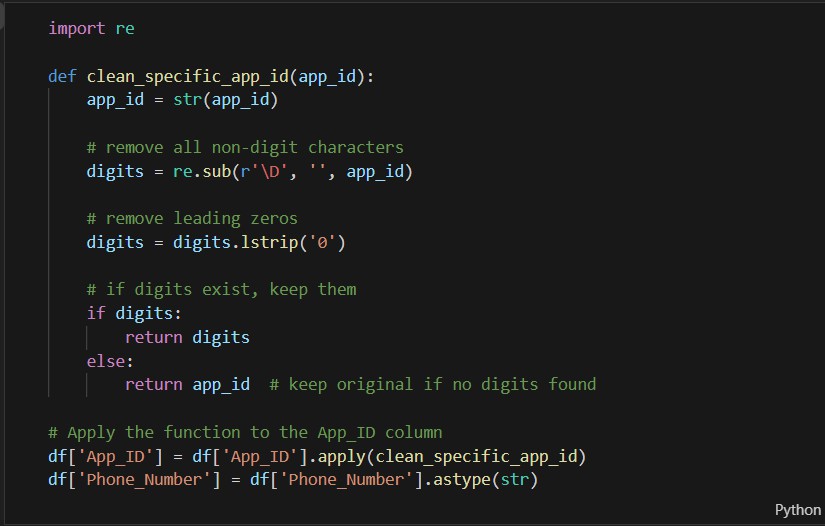
placed in the Phone\_Number column and Phone\_Number data is placed in the App\_ID column.

I tried to make all the App\_ID column data in one form, six digits format.

* + Swap the data of App\_ID and Phone\_Number if they are mixed up.

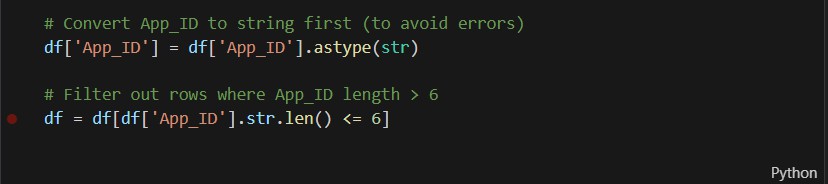


* + Remove all the non-digit characters from the APP\_ID
  + Remove A20 from the begining

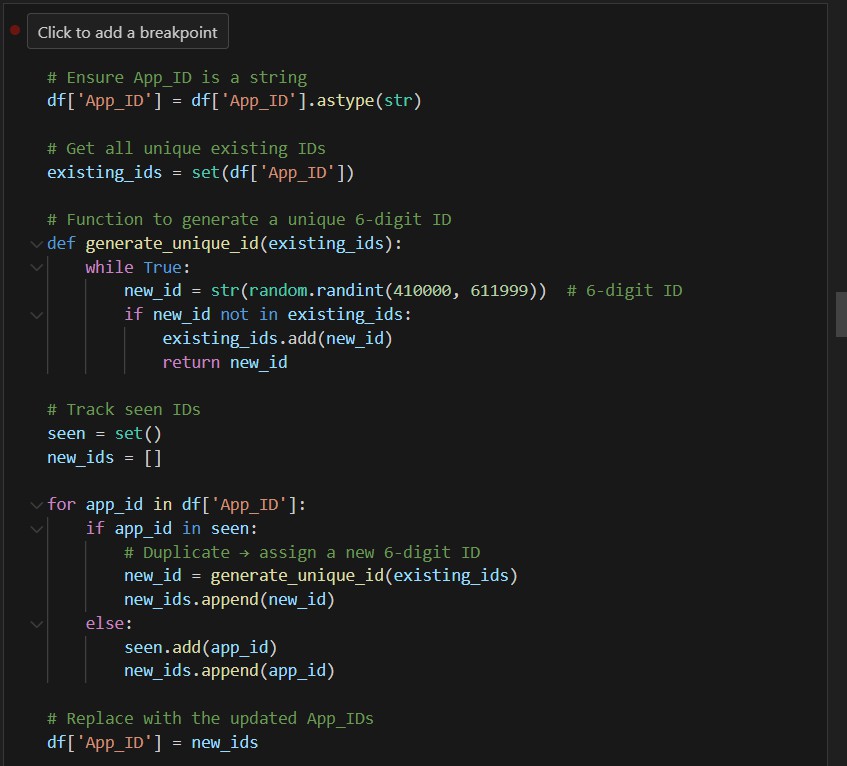




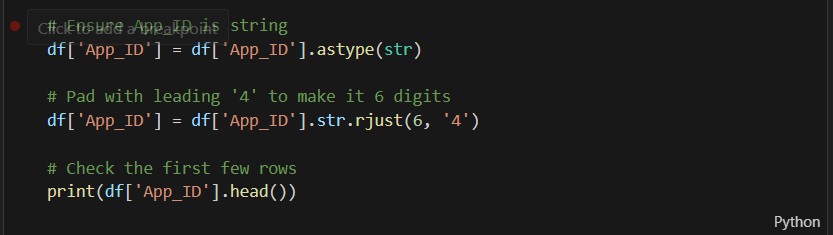
* + Remove those rows where there is still App\_ID with more than 6 digits.



* + Provide new ID if one ID is already existed.



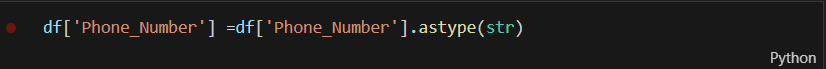
* + There is still remaining one with 5 digits. So I padded it with 4 to make this ID 6 digits.



Step 3: Cleaning the Phone\_Number Column

In the excel the some of the Country column is showing as a scientific number. Example, “2.52635E+11”.

So, to overcome this problem I just converted the data type into str. Previously it was objects type.

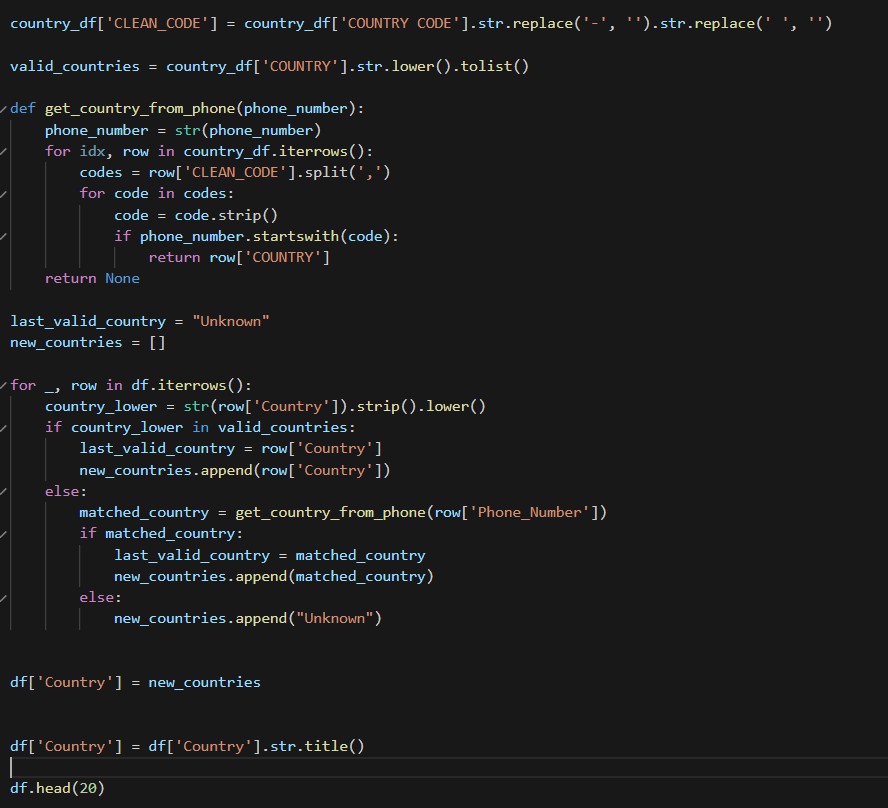


Step 4: Cleaning Country Column.

There are some where some email addresses and characters are pushed in the Country column. Example,

“[fnaeem1@hawk.iit.edu](mailto:fnaeem1@hawk.iit.edu)” “[masif3@hawk.iit.edu](mailto:masif3@hawk.iit.edu)”, “[ssundaram@hawk.iit.edu](mailto:ssundaram@hawk.iit.edu)”, “-”.

So, I tried to change this to country name with respect to Phone\_Number. All the country has unique country code to their phone number. I tried to match the country code of the phone number and put the valid country name in the Country Name Column.



**Data Quality Result summary:**

| **Operation** | **Detected / Before** | **Correction / After** |
| --- | --- | --- |
| Missing / invalid App\_ID | Some missing / invalid entries (',,,,', '/////////', 'na', 'naq', NaN) | Removed rows with missing / invalid App\_ID; dataset now contains only valid numeric App\_IDs |
| Duplicate rows | 16,489 duplicate rows | Removed duplicates; dataset shape reduced. |
| App\_ID column | 15,416 unique values, object type, some invalid | Converted to numeric to filter invalid values, removed invalid rows, then converted back to object; final valid unique App\_IDs = 15,160 |
| Country column | 150+ unique values with typos, lowercase, extra text, email-like strings | Corrected typos / capitalization; replaced email-like / multiple countries with 'Unknown'; standardized to valid country list |
| Phone\_Number column | 18,168 unique values, some with special chars, too short, or extremely long invalid numbers | Cleaned non-digit characters; replaced extremely long / invalid numbers in specific rows with 'Unknown'; valid numbers kept |
| University column | Only 1 unique value (Illinois Institute of Technology) | No change |
| Data types | Mostly object | Ensured App\_ID object, Phone\_Number object, Country object; datetime conversion applied if date column exists |
| Exported CSV | N/A | Exported as Cleaned\_ApplicantData.csv |

**Data Dictionary:**

| **Column Name** | **Corrected Data Type** | **Description of the Field** | **Notes on Any Changes from Original** |
| --- | --- | --- | --- |
| App\_ID | Object (TEXT) | 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 | Object (TEXT) | 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 with Power BI. |
| Country | Object (TEXT) | 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 | Object (TEXT) | Name of the applicant’s university | Original had only one unique value (Illinois Institute of Technology) → Stripped extra spaces for consistency; no value changes. |

**Dataset Overview after cleaning:**

* No null or missing values in any column.
* No duplicate rows.
* Correct datatypes for all columns.
* No inconsistencies in the data.
* Original valid data remained unchanged.
* Invalid entries were replaced without dropping unnecessary rows.

### ****Conclusion****

The Week-1 Data Visualization Trainee Early Internship deliverable has been successfully completed. The outreach, campaign, and applicant datasets have been systematically audited, cleaned, and documented. All key data quality issues, including missing values, duplicates, inconsistent formats, and invalid entries, were identified and addressed, resulting in structured and analysis-ready datasets. The Data Quality Report and Data Dictionary provide clear documentation of the cleaning process, ensuring transparency and usability for future analysis. This exercise has established a strong foundation for meaningful exploratory data analysis, visualization, and dashboard creation in the upcoming weeks.

**The Hyperlinks of our other documents**

1. **Our Cleaned Dataset(s) Links are below:**

<https://drive.google.com/drive/folders/1e7NsWi2JnRUtq7HwYhv6xvoq9ohe33wP?usp=sharing>

1. **Data Dictionary & Team Charter Links are below:**

<https://drive.google.com/drive/folders/1oW2O9aIP0tGNHM5cvm0Ta8vgJsC-pbWk?usp=sharing>

**Next Steps for Week 2**

1. Perform EDA on cleaned outreach, campaign, and application datasets.
2. Identify key trends, patterns, or anomalies.
3. Visualise insights with charts or tables.
4. Prepare a concise EDA Insights Report.
5. Design a dashboard showing main KPIs and segment breakdowns.
6. Ensure clear layout and narrative connecting visuals to insights.
7. Submit combined PDF with report and dashboard (file or screenshots).