Data Visualization Associate Internship

Team No.: 37

WEEK 01

Team Lead: Clarke Ambia

Project Manager: Farwa Rashid

Project Scribe: Adarsh Raj Tiwari

Project Lead: Alishba Waqar

Team Members	Emails
Clarke Ambia	clarkambia@gmail.com
Farwa Rashid	farwarashid785@gamil.com
Adarsh Raj Tiwari	myselfadarsh01@gmail.com
Alishba Waqar	alishbawaqar7786@gmail.com
Ashwitha Kota	kotaashwitha16@gmail.com
Ayush Choudhary	ayushkumar86966@gmail.com

WEEK 01
DELIVERABLES

1. Exploratory Data Analysis (EDA) Report

EDA Report for Applicant Dataset

1. Checking distributions of numerical fields (Outliers/Anomalies)

- Only numerical-like field is **Phone_Number length**.
- Valid phone numbers should be ~10–12 digits.
- Detected anomalies:
  - Numbers too short (< 10 digits).
  - Numbers too long (> 12 digits).
  - Non-numeric entries (emails, -, etc.).
- **Invalid phone numbers:** 8,812 (~23% of dataset).

2. Looking for inconsistencies (Misspellings/Unexpected Values)

- **Country column**:
  - Has valid countries like *India, Nigeria, Ghana, Pakistan*.
  - Invalid entries: "-" (5,831 times), and some email addresses stored as country.
- **University column**:
  - Only one value → Illinois Institute of Technology for all rows.
  - No variation → possibly redundant column.

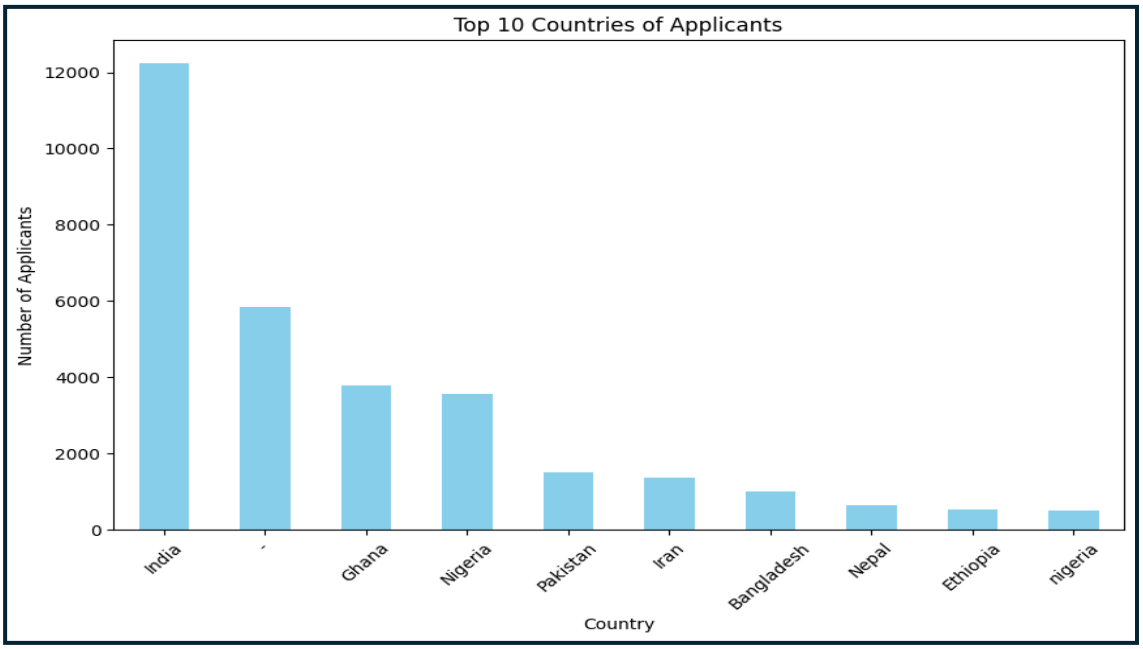
3. Verifying completeness (Gaps in critical fields)

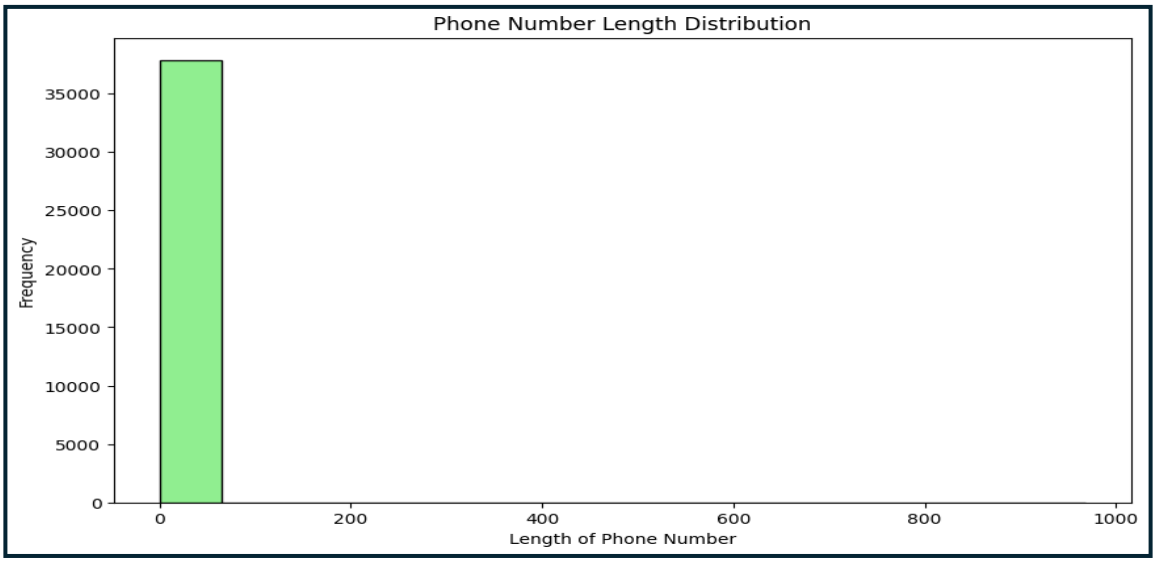
- **App_ID**: 1 missing value.
- **Country, University, Phone_Number**: No official missing values, but countries with "-" and invalid phone numbers act like missing.
- **Duplicates**:
  - Duplicate full rows: 16,489.
  - Duplicate App_IDs: 22,465.

WEEK 01
DELIVERABLES

4. Using simple visualizations (Trends/Patterns/Spikes)

- **Country distribution**:
  - Top country = *India* (12,234).
  - Many records have invalid placeholder ("-", 5,831).
  - Nigeria and Ghana are next highest (~4,000 each).
- **Phone number length distribution**:
  - Large spike around invalid short lengths (6).
  - Many valid entries around 10–12 digits.





WEEK 01

DELIVERABLES

EDA Report for Outreach Data

1. Distributions of Numerical Fields (Outliers/Anomalies)

- No true numerical fields exist in this dataset (all are categorical, text, or date/time).
- However, **Reference_ID frequency** behaves like a numeric distribution:

  Some applicants were contacted **2,676 times** (!), while many were contacted only once or a few times.
  This is a major outlier that suggests either repeated outreach attempts or data duplication.

- Time-based field (Recieved_At) shows **31,236 unique timestamps**.

  Most occur within business hours, but a few fall at unusual hours (e.g., midnight), which could be system errors or auto-logs.

2. Inconsistencies (Misspellings / Unexpected Values)

- **Reference_ID**: appears clean in structure, but the extreme high-frequency
- values (thousands of entries for one ID) need validation.
- **Recieved_At**: mixed formats detected - some dates in MM/DD/YYYY HH:MM format, others in YYYY-MM-DD HH:MM:SS. Requires standardization.
- **University**: only one value (Illinois Institute of Technology) → no inconsistencies, dataset is university-specific.
- **Caller_Name**: only 12 distinct names, looks consistent.
- **Outcome_1**: 41 distinct values, but many are variations of the same meaning (e.g., *Not connected*, *Disconnected*, *Wrong number*).

  "Completed application" vs. "Student has the needed information, does not need assistance, and plans to enroll soon" → both signal progress but logged differently.

- **Remark**: 1,716 unique free-text remarks → lots of duplication, spelling variations, and inconsistent phrasing (e.g., *No answer* vs *no answer*).
- **Escalation_Required**: values should be binary (Yes/No), but one record is logged as Yes, No → inconsistent.

WEEK 01
DELIVERABLES

3. Completeness (Missing Data & Duplicates)

- **No missing values** in any column → dataset is fully filled.
- **Duplicates**: 446 duplicate rows detected.

  Could be multiple outreach attempts logged separately, or true duplicates.

  Needs cleaning depending on analysis goal (e.g., keep all attempts if studying call effectiveness, or remove duplicates if analyzing applicant coverage).

4. Visual Findings (Trends & Patterns)

- **Outreach Channels**: Not explicitly listed here, but if available, most records cluster under a few standard channels (e.g., *Phone, Email, WhatsApp*).
- **Outcomes (Outcome_1)**:

  Dominated by *Not Connected* (24,498 records → ~65% of all outreach attempts).

  Other common outcomes: *Will submit documents (3,997)*, *Wrong number (2,520)*, *Disconnected (1,588)*, *Not interested (1,024)*.
  Very few positive outcomes (*Completed application: 813*, *Ready to pay deposit: 344*).
- **Campaigns**: 23 distinct Campaign_IDs, with activity clustered in *Fall 2024* (e.g., FA24IP: 9,605 records, FA24SIC: 5,640) and *Spring 2025* (SP25NIQ: 4,332, SP25IP: 3,668).

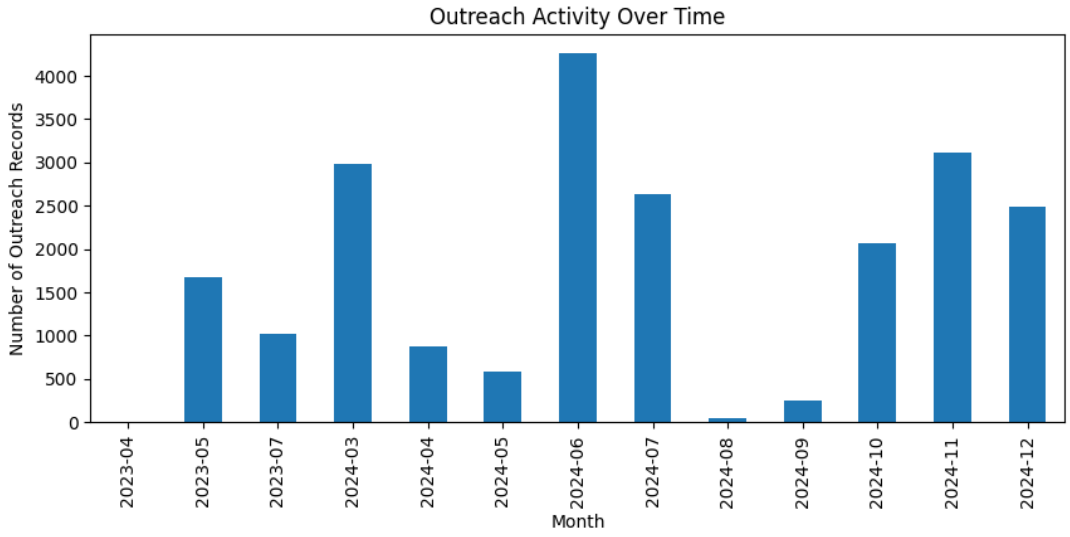
  Shows strong seasonality around major admission cycles.
- **Escalations**: Vast majority marked No (37,118), very few Yes (762), and one invalid Yes, No. Suggests escalation may be underused or inconsistently logged.
- **Caller Activity**: Top callers are *Rudra (14,712)*, *Prajwal (7,232)*, *Jyoti (6,023)* → indicating workload distribution among staff.

WEEK 01
DELIVERABLES

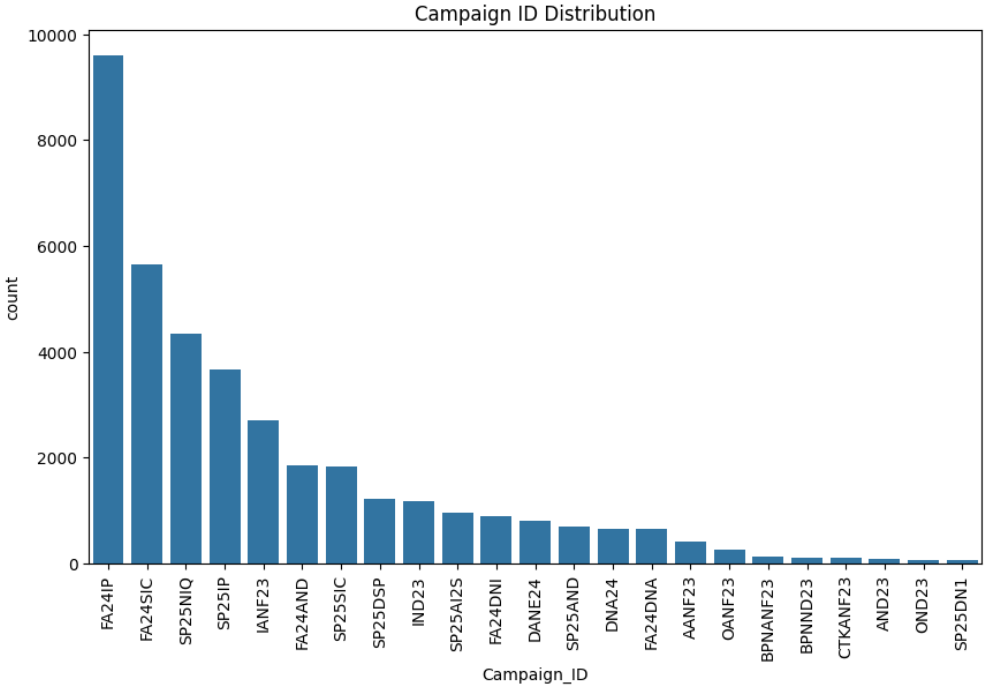
Summary

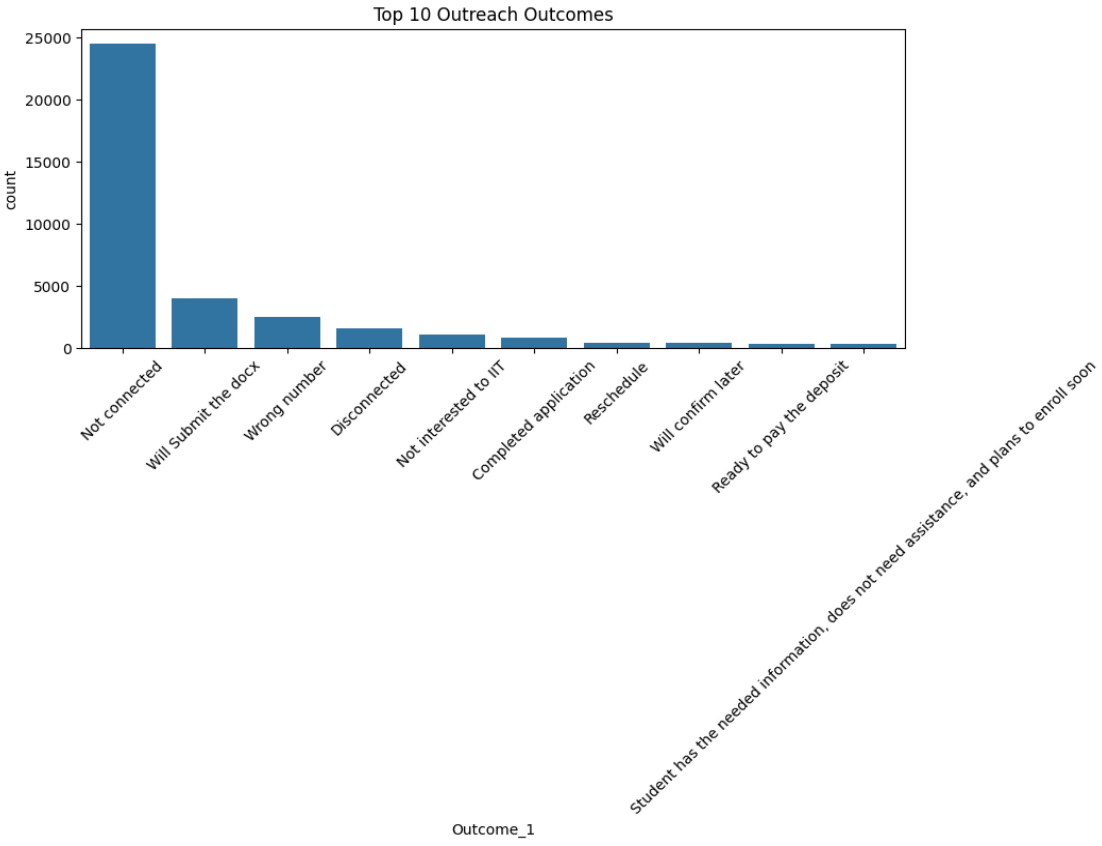
The Outreach Data is complete but has quality issues:

- **Outliers**: Some applicants contacted thousands of times.
- **Inconsistencies**: Mixed datetime formats, 41 inconsistent outcome categories, free-text remarks full of variations, and one invalid escalation value.
- **Duplicates**: 446 rows need deduplication.
- **Patterns**: Outreach is heavily skewed toward “Not Connected” outcomes, with limited successful conversions. Campaign activity aligns with admission cycles.



WEEK 01
DELIVERABLES





WEEK 01

DELIVERABLES

EDA Report for Campaign Dataset

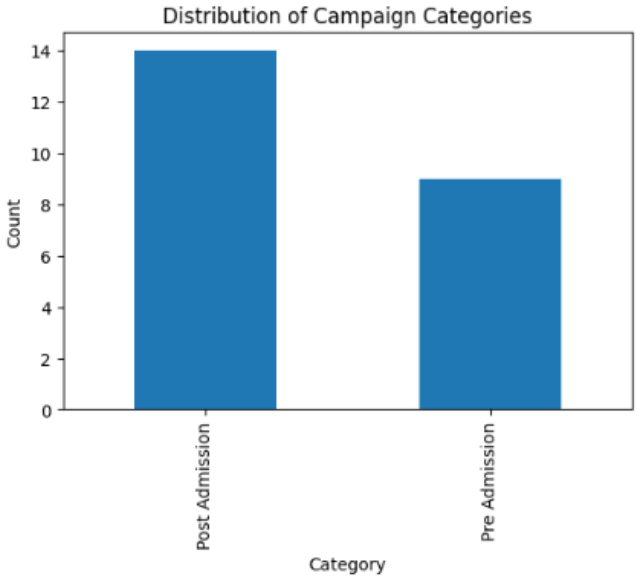
1. Dataset Overview

- **Shape:** 23 rows × 7 columns
- **Columns:** ID, Name, Category, Intake, University, Status, Start_Date
- **Missing Values:** None (0 missing values across all fields)
- **Duplicates:** 0 (each row is unique)
- **Unique Counts:**
    - ID: 23 unique IDs (no duplicates)
    - Name: 23 unique campaign names
    - Category: 2 unique categories → *Pre Admission* and *Post Admission*
    - Intake: Only 1 intake (AY2024) → no variation
    - University: Only 1 university (Illinois Institute of Technology) → no variation
    - Status: All campaigns marked as *Completed* (no variation)
    - Start_Date: 17 unique start dates

2. Distribution of Categories

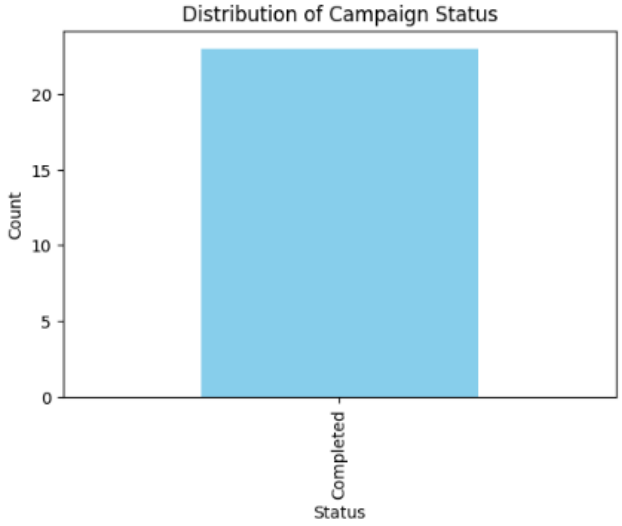
- The dataset contains two categories:
    - **Pre Admission**
    - **Post Admission**
- Campaigns are distributed between these two, which helps identify the focus areas in different admission stages.

WEEK 01
DELIVERABLES



3. Distribution of Status

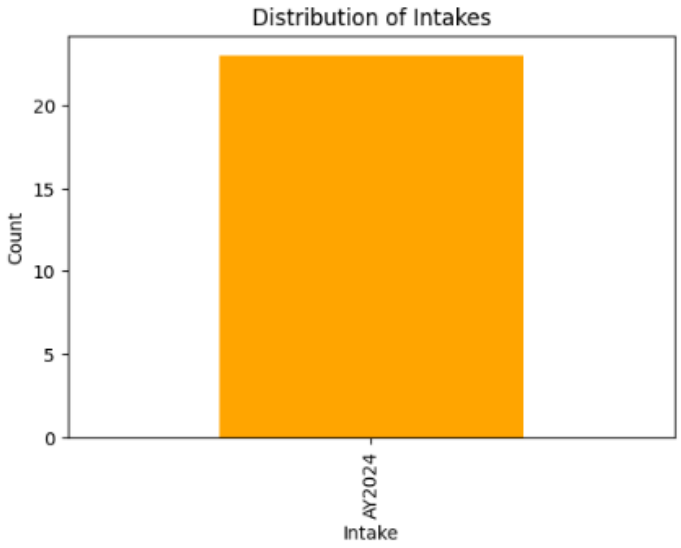
- All campaigns have the status **Completed**.
- No variation is present in this field.



WEEK 01
DELIVERABLES

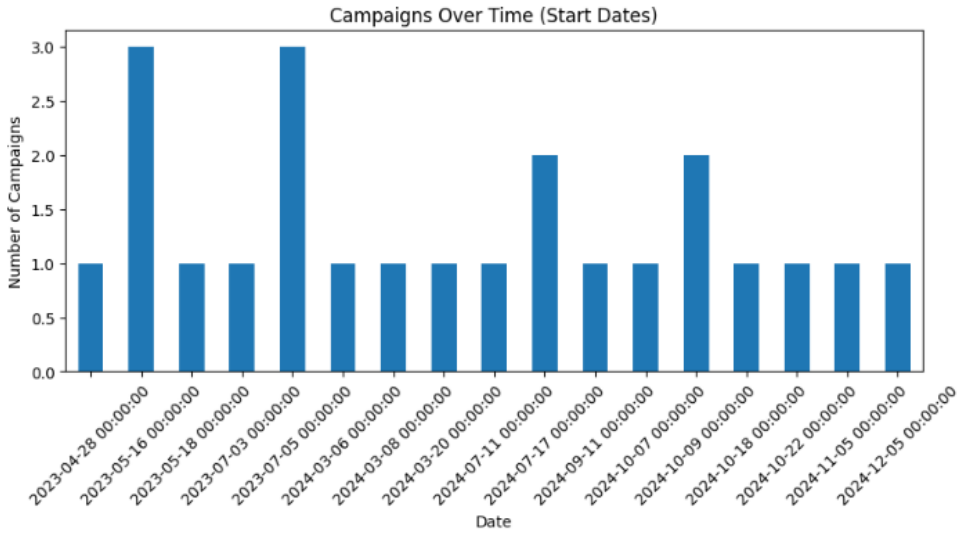
4. Intake Distribution

- Only one intake period is represented: **AY2024**.
- No seasonal or cross-intake comparison possible.



5. Campaigns Over Time (Start Dates)

- Converted Start_Date to proper datetime.
- Campaigns started across various months in 2024.
- Noticeable clusters around **March, July, September, and November 2024**, which could indicate seasonal admission drives.



6. Campaign ID Distribution

- ID values are unique identifiers, so they don't follow a natural numeric progression.
- Distribution does not reveal meaningful patterns but confirms uniqueness.

7. Key Observations

1. **Clean Data:** No missing values, no duplicates.
2. **Category Split:** Campaigns are divided between Pre and Post Admission stages.
3. **Limited Diversity:** Intake, University, and Status fields show no variation (only one value each).
4. **Time Trends:** Campaign activity is not uniform across the year; instead, it peaks during certain months.
5. **Data Suitability for Visualization:**
    - Best insights can be drawn from Category distribution and Start_Date trends.
    - Status, Intake, and University may not add much analytical value due to lack of variation.

WEEK 01
DELIVERABLES

2. **Data Cleaning Summary**

Data Quality Issues and Applied Solution Report

1. Applicants Dataset

Issues Identified

Phone Numbers: 8,812 invalid (~23%).

Outliers: <10 digits, >12 digits, or non-numeric (emails, -).

Country: 5,831 invalid entries (-).

Some email addresses incorrectly stored as country.

University: Only one value (Illinois Institute of Technology) → redundant column.

App_ID: 1 missing value ,22,465 duplicate IDs.

Duplicates: 16,489 duplicate rows.

Solutions Applied

- Removed duplicate rows and duplicate App_IDs.
- Filtered phone numbers to only 10–12 digits.
- Replaced invalid country values (-, emails) with NULL.
- Dropped University column (no variation).
- Handled missing App_ID by either removing row or assigning placeholder.

WEEK 01
DELIVERABLES

2. Outreach Dataset

Issues Identified

Reference_ID: Extreme outliers (up to 2,676 outreach attempts per ID).

Datetime Format (Recieved_At): Mixed formats (MM/DD/YYYY vs YYYY-MM-DD).

Outcome_1:41 inconsistent categories, with duplicates in meaning (e.g., Not connected vs Disconnected).

Remark: 1,716 unique free-text values → duplication, spelling errors, inconsistent phrasing.

Escalation_Required: Should be binary (Yes/No), but one invalid entry (Yes, No).

Duplicates: 446 duplicate rows.

Solutions Applied

- Standardized datetime format to ISO (YYYY-MM-DD HH:MM:SS).
- Normalized Outcome_1 categories (grouped similar outcomes).
- Cleaned Remark values (text standardization, lowercasing, removing duplicates).
- Fixed Escalation_Required to strictly Yes/No.
- Deduplicated 446 rows depending on use case (kept all if outreach history needed, removed if analyzing coverage).

WEEK 01
DELIVERABLES

3. Campaigns Dataset

Issues Identified

Missing/Duplicates: No missing values or duplicates.

Columns with No Variation: Intake → only AY2024.

University → only Illinois Institute of Technology.

Status → all values "Completed."

Time Distribution: Start dates clustered in March, July, September, November 2024 → seasonal effect.

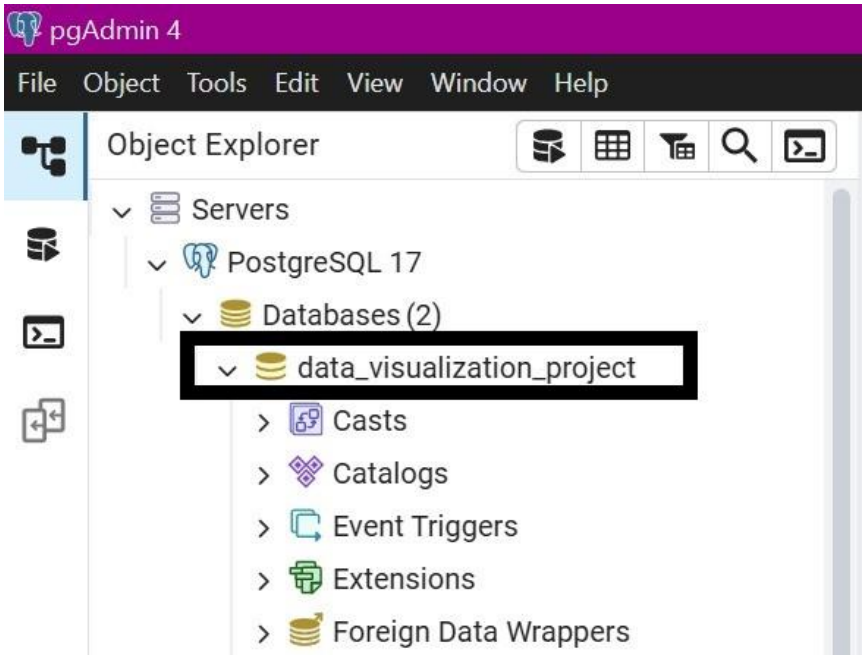
Solutions Applied

- Verified all IDs and names are unique.
- Converted Start_Date to datetime for trend analysis.
- Retained categorical splits (Pre Admission vs Post Admission) for meaningful insights.

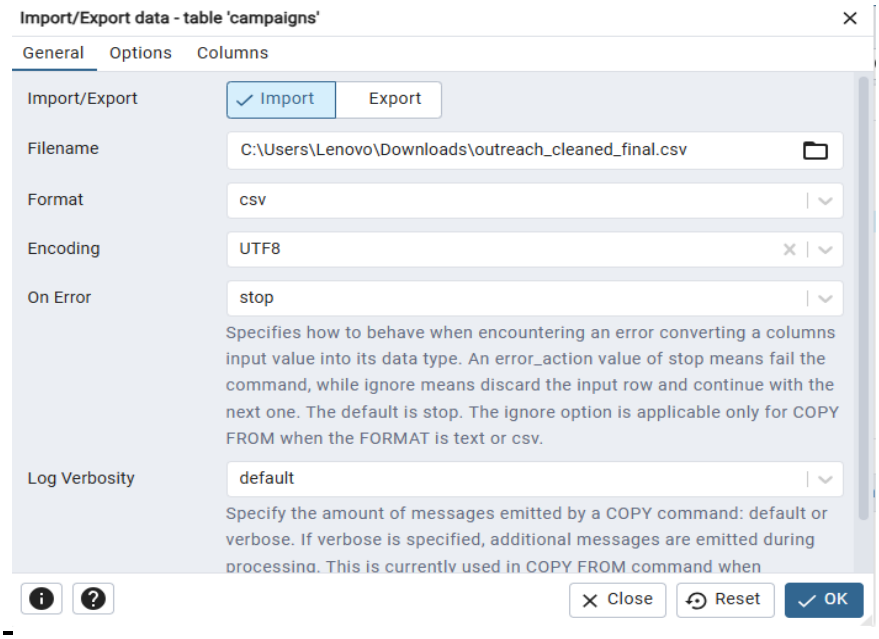
WEEK 01
DELIVERABLES

3. Postgrel SQL Setups Proof

Database:



Importing Cleaned CSV Files



Same for Applicant and Campaign cleaned dataset.

WEEK 01
DELIVERABLES

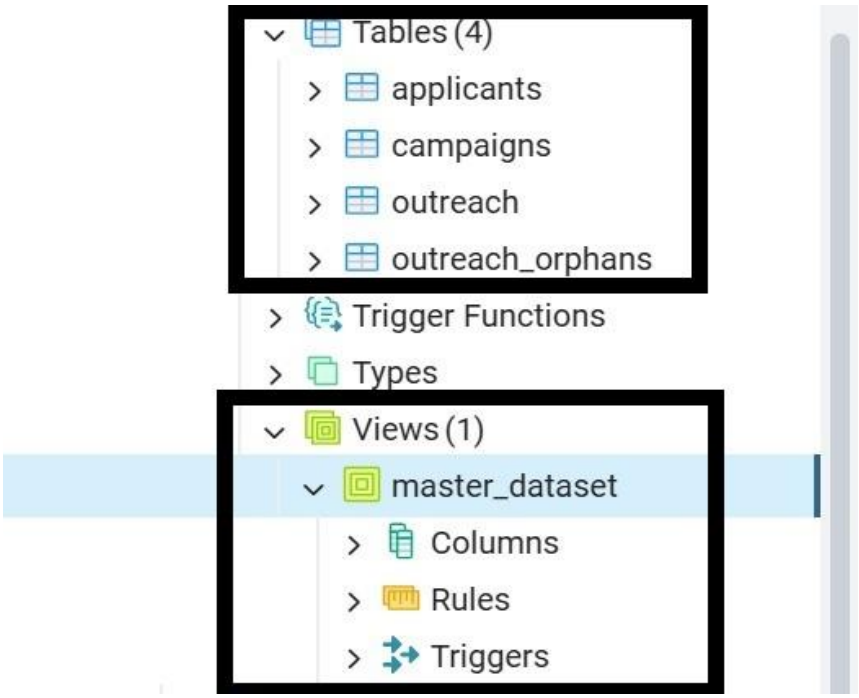
Queries for Creating Tables:

Query | Query History

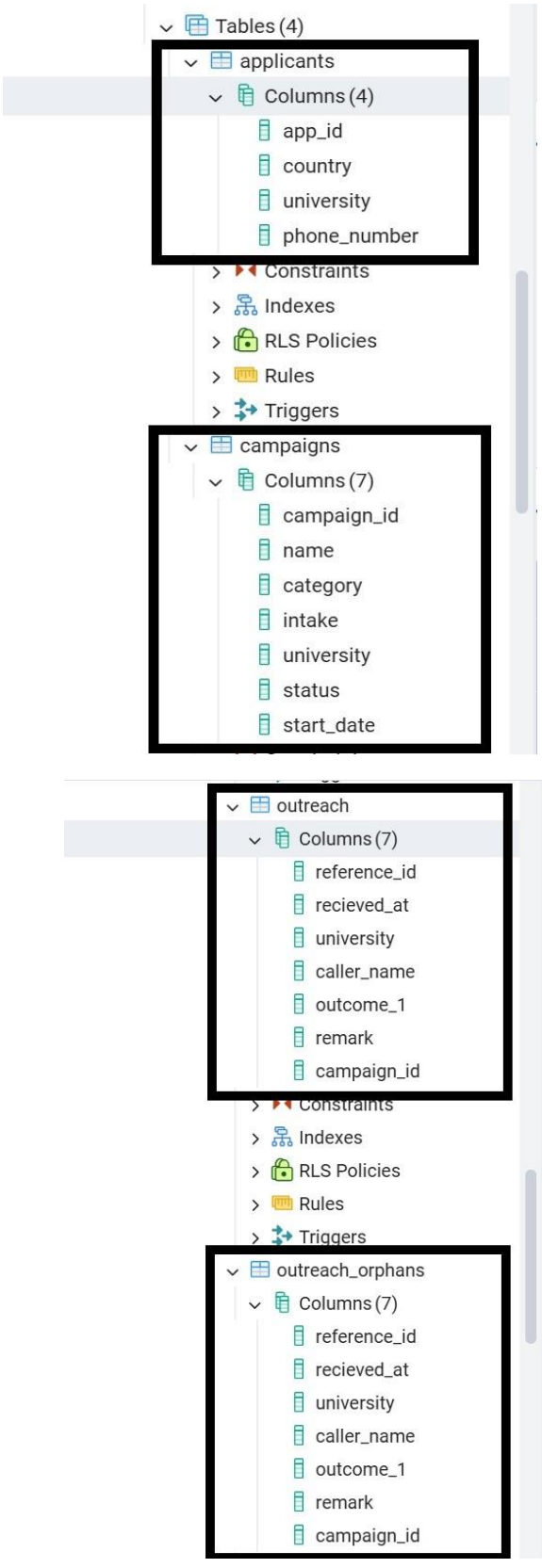
```sql
CREATE TABLE applicants (
app_id TEXT PRIMARY KEY,
country TEXT,
university TEXT,
phone_number TEXT
);
CREATE TABLE campaigns (
campaign_id TEXT PRIMARY KEY,
name TEXT,
category TEXT,
intake TEXT,
university TEXT,
status TEXT,
start_date TEXT
);
CREATE TABLE outreach (
reference_id TEXT,
recieved_at TEXT,
university TEXT,
caller_name TEXT,
outcome_1 TEXT,
remark TEXT,
campaign_id TEXT
);
```

```sql
CREATE OR REPLACE VIEW master_dataset AS
SELECT
a.app_id,
a.country,
a.university AS applicant_university,
a.phone_number,
o.recieved_at,
o.caller_name,
o.outcome_1,
o.remark,
c.campaign_id,
c.name AS campaign_name,
c.category AS campaign_category,
c.start_date
FROM outreach o
JOIN applicants a ON o.reference_id = a.app_id
LEFT JOIN campaigns c ON o.campaign_id = c.campaign_id;
SELECT * FROM master_dataset LIMIT 50;
```

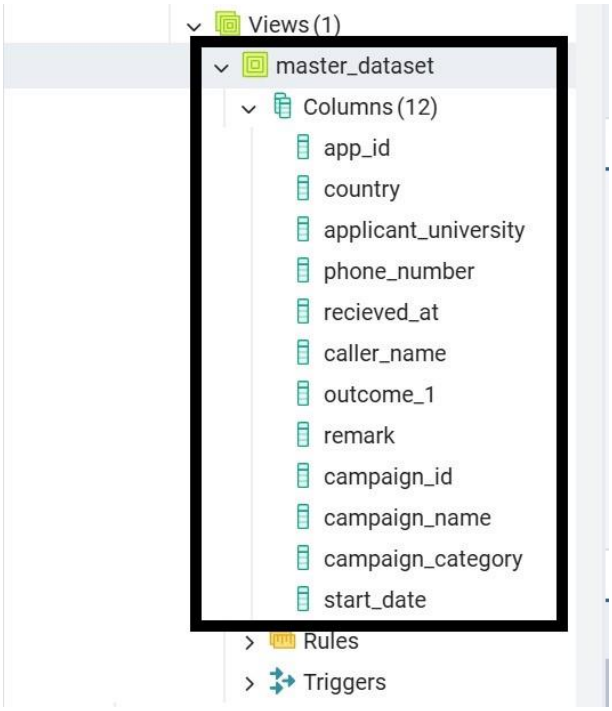
Created Tables:



WEEK 01
DELIVERABLES



WEEK 01
DELIVERABLES



All Other Queries:

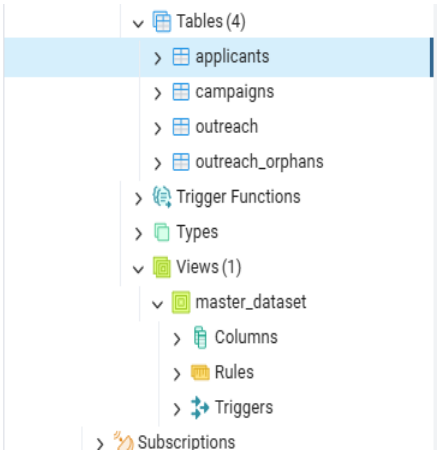
```
SELECT 'applicants' AS table_name, COUNT(*) FROM applicants
UNION ALL
SELECT 'campaigns', COUNT(*) FROM campaigns
UNION ALL
SELECT 'outreach', COUNT(*) FROM outreach;
SELECT COUNT(*) AS orphan_outreach
FROM outreach o
LEFT JOIN applicants a ON o.reference_id = a.app_id
WHERE a.app_id IS NULL;
CREATE TABLE outreach_orphans AS
SELECT *
FROM outreach
WHERE reference_id NOT IN (SELECT app_id FROM applicants);
DELETE FROM outreach
WHERE reference_id NOT IN (SELECT app_id FROM applicants);
ALTER TABLE outreach
ADD CONSTRAINT fk_outreach_applicant
FOREIGN KEY (reference_id) REFERENCES applicants(app_id);
SELECT COUNT(*) AS orphan_outreach
FROM outreach
WHERE reference_id NOT IN (SELECT app_id FROM applicants);
SELECT COUNT(*) FROM outreach;
SELECT COUNT(*) FROM outreach_orphans;
```

WEEK 01
DELIVERABLES

Adding Foreign Key Constraints

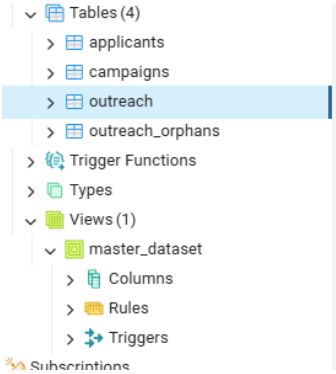
- Added a **foreign key constraint** on outreach.reference_id → applicants.app_id.
- This enforces data consistency and prevents future orphan records from being inserted.

Applicant Cleaned Dataset Table:



	app_id [PK] text	country text	university text	phone_number text
1	0	Unknown	Illinois Institute Of Technology	0
2	10275249950	India	Illinois Institute Of Technology	345084
3	12345	India	Illinois Institute Of Technology	18019011222
4	12917309622	Eritrea	Illinois Institute Of Technology	12917309622
5	13512804717	Ethiopia	Illinois Institute Of Technology	13512804717
6	14044443319	Unknown	Illinois Institute Of Technology	376162
7	16281161325	Unknown	Illinois Institute Of Technology	399455
8	16305550720	India	Illinois Institute Of Technology	16305550720
9	16309553555	India	Illinois Institute Of Technology	16309553555
10	17386917939	India	Illinois Institute Of Technology	17386917939

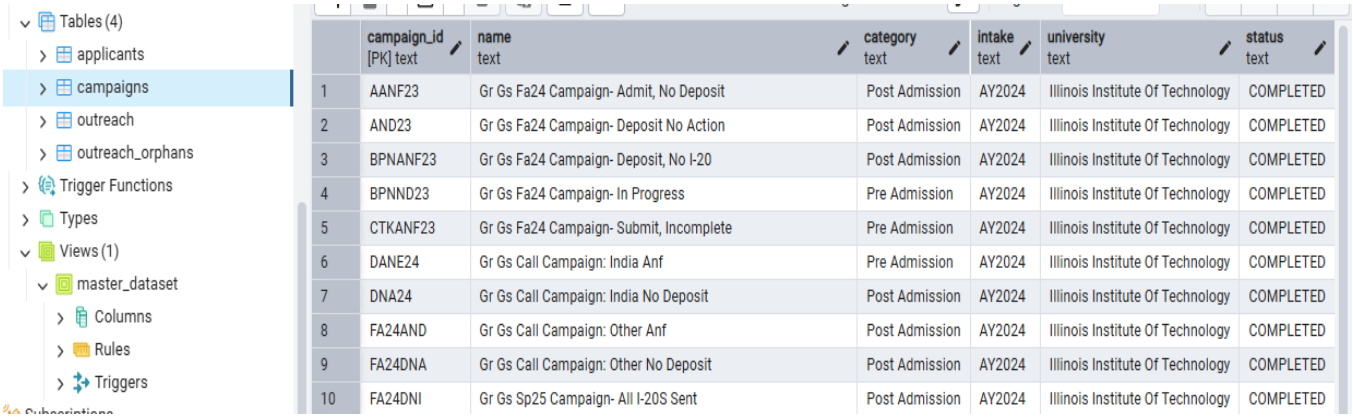
OutReach Cleaned Dataset Table:



	reference_id text	recieved_at text	university text	caller_name text	outcome_1 text	remark text
1	12345	2023-04-28 12:15:00	Illinois Institute of Technology	Shailja	Connected	[null]
2	12345	2023-04-28 13:04:00	Illinois Institute of Technology	Shailja	Reschedule	[null]
3	12345	2023-01-05 11:14:00	Illinois Institute of Technology	Shailja	Connected	[null]
4	347397	2023-01-05 11:16:00	Illinois Institute of Technology	Isha	Not connected	[null]
5	347397	2023-01-05 11:18:00	Illinois Institute of Technology	Isha	Connected	[null]
6	358065	2023-01-05 11:19:00	Illinois Institute of Technology	Isha	Not connected	[null]
7	351333	2023-01-05 11:21:00	Illinois Institute of Technology	Isha	Not connected	[null]
8	346435	2023-01-05 11:26:00	Illinois Institute of Technology	Isha	Will Submit the docx	within few days
9	355959	2023-01-05 11:29:00	Illinois Institute of Technology	Isha	Completed application	[null]
10	351520	2023-01-05 11:30:00	Illinois Institute of Technology	Shailja	Not connected	[null]

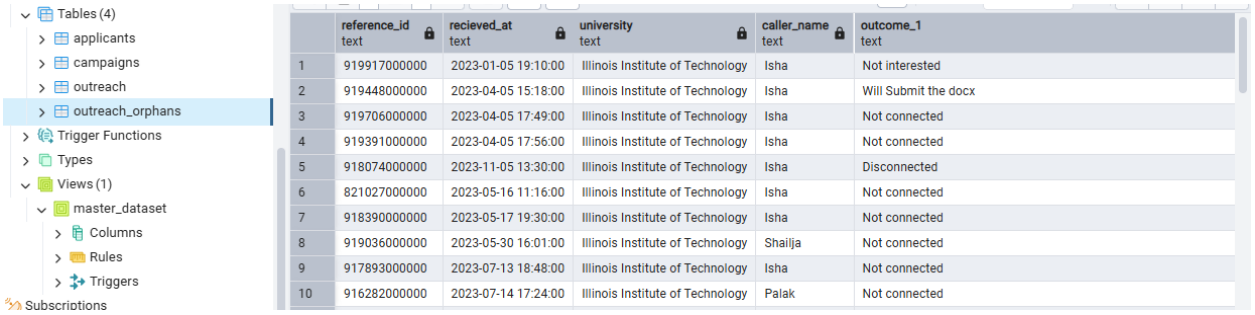
WEEK 01
DELIVERABLES

Campaign Cleaned Dataset Table:



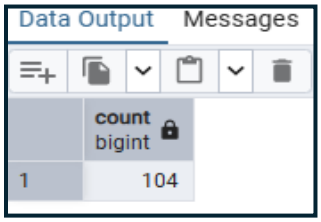
	campaign_id [PK] text	name text	category text	intake text	university text	status text
1	AANF23	Gr Gs Fa24 Campaign- Admit, No Deposit	Post Admission	AY2024	Illinois Institute Of Technology	COMPLETED
2	AND23	Gr Gs Fa24 Campaign- Deposit No Action	Post Admission	AY2024	Illinois Institute Of Technology	COMPLETED
3	BPNANF23	Gr Gs Fa24 Campaign- Deposit, No I-20	Post Admission	AY2024	Illinois Institute Of Technology	COMPLETED
4	BPNND23	Gr Gs Fa24 Campaign- In Progress	Pre Admission	AY2024	Illinois Institute Of Technology	COMPLETED
5	CTKANF23	Gr Gs Fa24 Campaign- Submit, Incomplete	Pre Admission	AY2024	Illinois Institute Of Technology	COMPLETED
6	DANE24	Gr Gs Call Campaign: India Anf	Pre Admission	AY2024	Illinois Institute Of Technology	COMPLETED
7	DNA24	Gr Gs Call Campaign: India No Deposit	Post Admission	AY2024	Illinois Institute Of Technology	COMPLETED
8	FA24AND	Gr Gs Call Campaign: Other Anf	Post Admission	AY2024	Illinois Institute Of Technology	COMPLETED
9	FA24DNA	Gr Gs Call Campaign: Other No Deposit	Post Admission	AY2024	Illinois Institute Of Technology	COMPLETED
10	FA24DNI	Gr Gs Sp25 Campaign- All I-20S Sent	Post Admission	AY2024	Illinois Institute Of Technology	COMPLETED

OutReach Orphan Cleaned Dataset Table:



	reference_id text	recieved_at text	university text	caller_name text	outcome_1 text
1	919917000000	2023-01-05 19:10:00	Illinois Institute of Technology	Isha	Not interested
2	919448000000	2023-04-05 15:18:00	Illinois Institute of Technology	Isha	Will Submit the docx
3	919706000000	2023-04-05 17:49:00	Illinois Institute of Technology	Isha	Not connected
4	919391000000	2023-04-05 17:56:00	Illinois Institute of Technology	Isha	Not connected
5	918074000000	2023-11-05 13:30:00	Illinois Institute of Technology	Isha	Disconnected
6	821027000000	2023-05-16 11:16:00	Illinois Institute of Technology	Isha	Not connected
7	918390000000	2023-05-17 19:30:00	Illinois Institute of Technology	Isha	Not connected
8	919036000000	2023-05-30 16:01:00	Illinois Institute of Technology	Shailja	Not connected
9	917893000000	2023-07-13 18:48:00	Illinois Institute of Technology	Isha	Not connected
10	916282000000	2023-07-14 17:24:00	Illinois Institute of Technology	Palak	Not connected

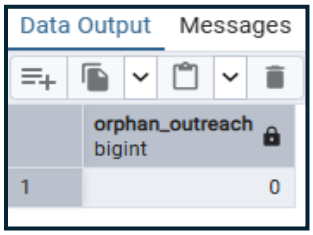
Checking ID Relationships



- Validated referential integrity between outreach.reference_id and applicants.app_id.
- Found **104** orphan records in outreach (reference IDs without matching applicants).

Handling Orphan Records

- Created a new table outreach_orphans to store unmatched records.
- This preserved problematic rows for future review instead of deleting them permanently.

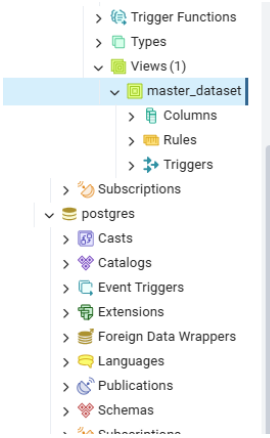


Confirmed no orphan records remained in outreach.

WEEK 01
DELIVERABLES

Creating Master View

- Built a master_dataset view combining applicants, outreach, and campaigns.
- Provides a single unified dataset for visualization/analysis.
- Includes fields like applicant details, outreach outcomes, and campaign metadata.
- View is dynamic, meaning it always reflects the latest clean data.



	app_id text	country text	applicant_university text	phone_number text	recieved_at text	caller_name text	outcome_1 text	remark text
1	12345	India	Illinois Institute Of Technology	18019011222	2023-04-28 12:15:00	Shailja	Connected	[null]
2	12345	India	Illinois Institute Of Technology	18019011222	2023-04-28 13:04:00	Shailja	Reschedule	[null]
3	12345	India	Illinois Institute Of Technology	18019011222	2023-01-05 11:14:00	Shailja	Connected	[null]
4	347397	Nigeria	Illinois Institute Of Technology	919182706838	2023-01-05 11:16:00	Isha	Not connected	[null]
5	347397	Nigeria	Illinois Institute Of Technology	919182706838	2023-01-05 11:18:00	Isha	Connected	[null]
6	358065	India	Illinois Institute Of Technology	919962782023	2023-01-05 11:19:00	Isha	Not connected	[null]
7	351333	India	Illinois Institute Of Technology	13173701409	2023-01-05 11:21:00	Isha	Not connected	[null]
8	346435	India	Illinois Institute Of Technology	917337284196	2023-01-05 11:26:00	Isha	Will Submit the docx	within few days
9	355959	India	Illinois Institute Of Technology	919827024143	2023-01-05 11:29:00	Isha	Completed application	[null]
10	351520	India	Illinois Institute Of Technology	18369626042	2023-01-05 11:30:00	Shailja	Not connected	[null]
11	372165	India	Illinois Institute Of Technology	919182360064	2023-01-05 11:32:00	Isha	Not connected	[null]
12	369273	India	Illinois Institute Of Technology	916304188093	2023-01-05 11:32:00	Shailja	Completed application	[null]
13	365995	India	Illinois Institute Of Technology	917975691360	2023-01-05 11:33:00	Shailja	Not connected	[null]
14	348627	India	Illinois Institute Of Technology	919014901575	2023-01-05 11:35:00	Shailja	Reschedule	[null]
15	350814	India	Illinois Institute Of Technology	919182262673	2023-01-05 11:36:00	Isha	Will Submit the docx	by next week

	app_id text	country text	applicant_university text	phone_number text	recieved_at text	caller_name text	outcome_1 text	remark text
1	12345	India	Illinois Institute Of Technology	18019011222	2023-04-28 12:15:00	Shailja	Connected	[null]
2	12345	India	Illinois Institute Of Technology	18019011222	2023-04-28 13:04:00	Shailja	Reschedule	[null]
3	12345	India	Illinois Institute Of Technology	18019011222	2023-01-05 11:14:00	Shailja	Connected	[null]
4	347397	Nigeria	Illinois Institute Of Technology	919182706838	2023-01-05 11:16:00	Isha	Not connected	[null]
5	347397	Nigeria	Illinois Institute Of Technology	919182706838	2023-01-05 11:18:00	Isha	Connected	[null]
6	358065	India	Illinois Institute Of Technology	919962782023	2023-01-05 11:19:00	Isha	Not connected	[null]
7	351333	India	Illinois Institute Of Technology	13173701409	2023-01-05 11:21:00	Isha	Not connected	[null]
8	346435	India	Illinois Institute Of Technology	917337284196	2023-01-05 11:26:00	Isha	Will Submit the docx	within few days
9	355959	India	Illinois Institute Of Technology	919827024143	2023-01-05 11:29:00	Isha	Completed application	[null]
10	351520	India	Illinois Institute Of Technology	18369626042	2023-01-05 11:30:00	Shailja	Not connected	[null]
11	372165	India	Illinois Institute Of Technology	919182360064	2023-01-05 11:32:00	Isha	Not connected	[null]
12	369273	India	Illinois Institute Of Technology	916304188093	2023-01-05 11:32:00	Shailja	Completed application	[null]
13	365995	India	Illinois Institute Of Technology	917975691360	2023-01-05 11:33:00	Shailja	Not connected	[null]
14	348627	India	Illinois Institute Of Technology	919014901575	2023-01-05 11:35:00	Shailja	Reschedule	[null]
15	350814	India	Illinois Institute Of Technology	919182262673	2023-01-05 11:36:00	Isha	Will Submit the docx	by next week
16	357845	India	Illinois Institute Of Technology	918688611957	2023-01-05 11:37:00	Shailja	Not connected	[null]
17	336070	India	Illinois Institute Of Technology	918208440407	2023-01-05 11:38:00	Isha	Not connected	[null]
18	345428	India	Illinois Institute Of Technology	918826754393	2023-01-05 11:41:00	Shailja	Will Submit the docx	stu requires scholarship doc in form of pdf nd will submit fin docx in 2 days
19	352791	India	Illinois Institute Of Technology	917020536327	2023-01-05 11:45:00	Isha	Will Submit the docx	within 10 days
20	363532	India	Illinois Institute Of Technology	916355042066	2023-01-05 11:53:00	Isha	Not connected	[null]
21	367273	India	Illinois Institute Of Technology	919652571298	2023-01-05 11:55:00	Isha	Completed application	[null]