EDA & Data Validation Report

PRESENTED BY ---

FARHAN HUSSAIN
M. NOUMAN ARSHAD
SHIVANSH RAJPUT
DIA AGARWAL
YUVA MIRSHA
AJAY KUMAR KASAUDHAN
NAHEED AKHTAR
SAURAV KUMAR
MRUNALI UMAK
JAYDEEP HINGU

A comprehensive analysis of data integrity and quality

Agenda: Key Topics for Today's Presentation

01

Overview of data sources and their significance in the analysis process.

02

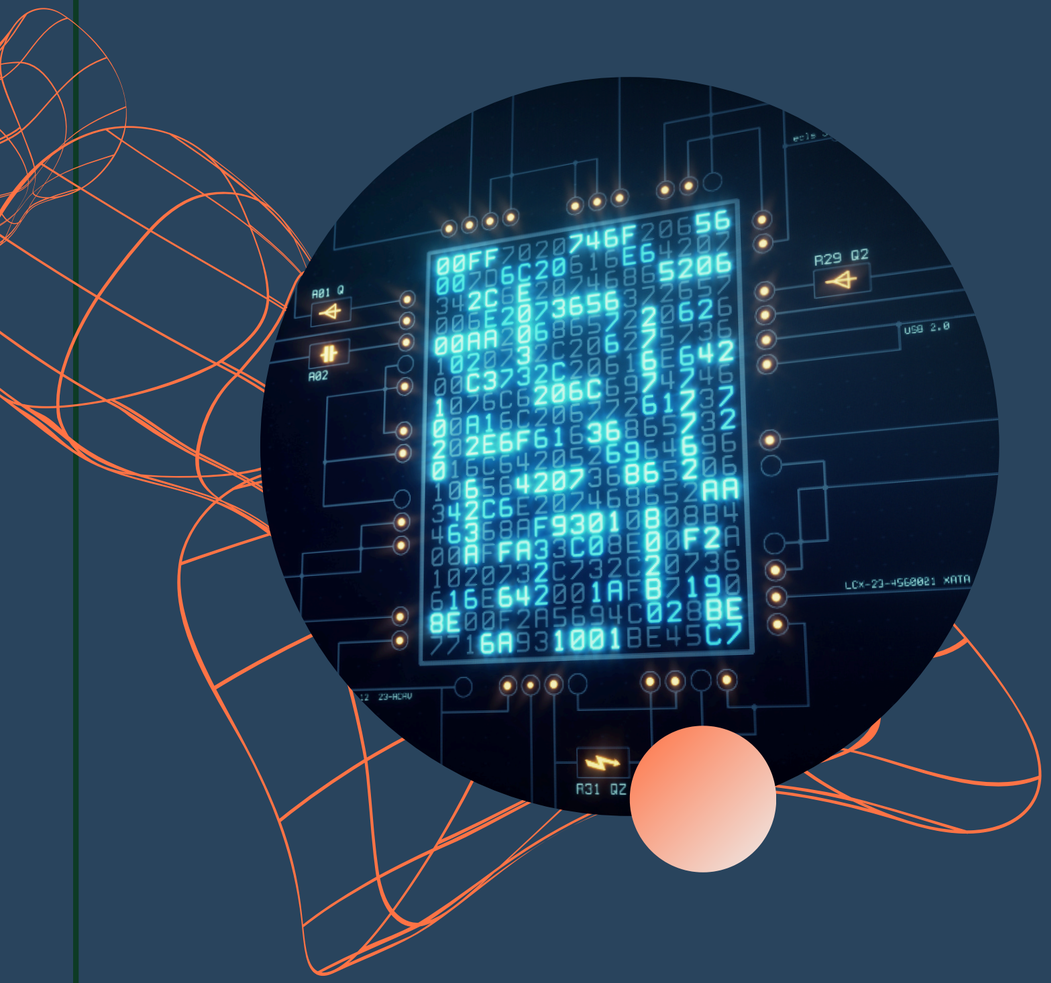
Exploring EDA steps taken for data cleaning and preparation.

03

Detailed description of the data validation procedures implemented.

04

Comparison of results before and after the data cleaning efforts.



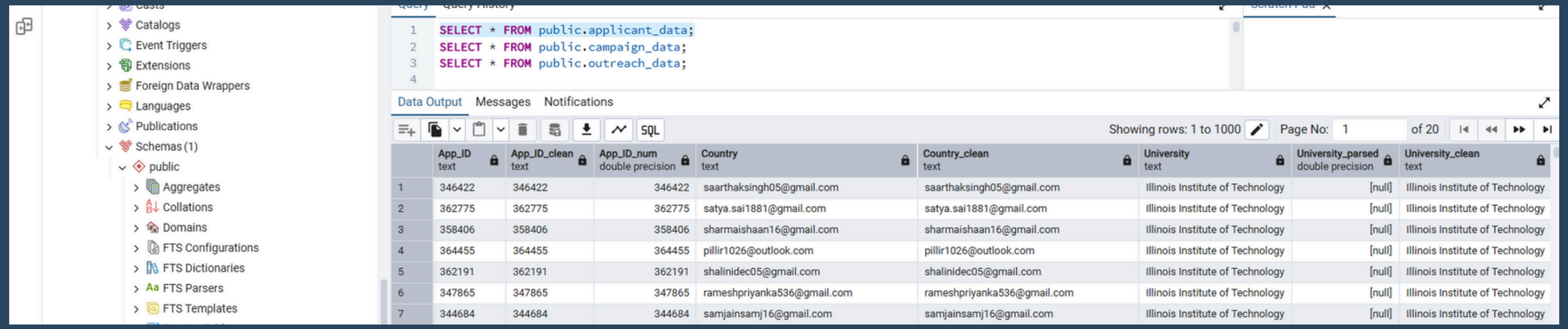
Data Sources Overview

Consolidating diverse datasets for analysis

The dataset comprises **Applicant Data**, **Outreach Data**, and **Campaign Data**, which are consolidated into a Master Table of **~24 lakh rows**. This integration ensures a comprehensive view for effective analysis and decision-making based on actionable insights.

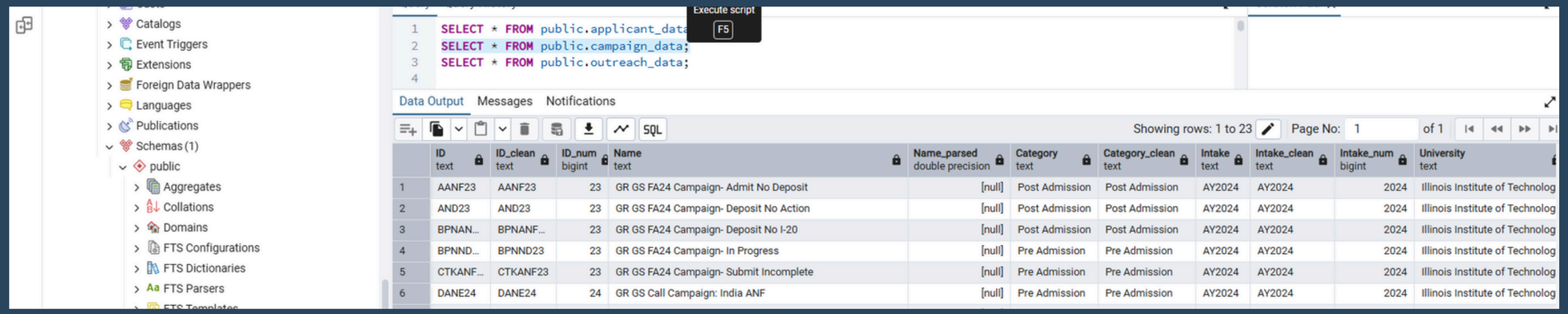
Data Set Overview

Applicant_Data



	App_ID text	App_ID_clean text	App_ID_num double precision	Country text	Country_clean text	University text	University_parsed double precision	University_clean text
1	346422	346422	346422	saarthaksingh05@gmail.com	saarthaksingh05@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
2	362775	362775	362775	satya.sai1881@gmail.com	satya.sai1881@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
3	358406	358406	358406	sharmaishaan16@gmail.com	sharmaishaan16@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
4	364455	364455	364455	pillir1026@outlook.com	pillir1026@outlook.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
5	362191	362191	362191	shalinidec05@gmail.com	shalinidec05@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
6	347865	347865	347865	rameshpriyanka536@gmail.com	rameshpriyanka536@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
7	344684	344684	344684	samjainsamj16@gmail.com	samjainsamj16@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology

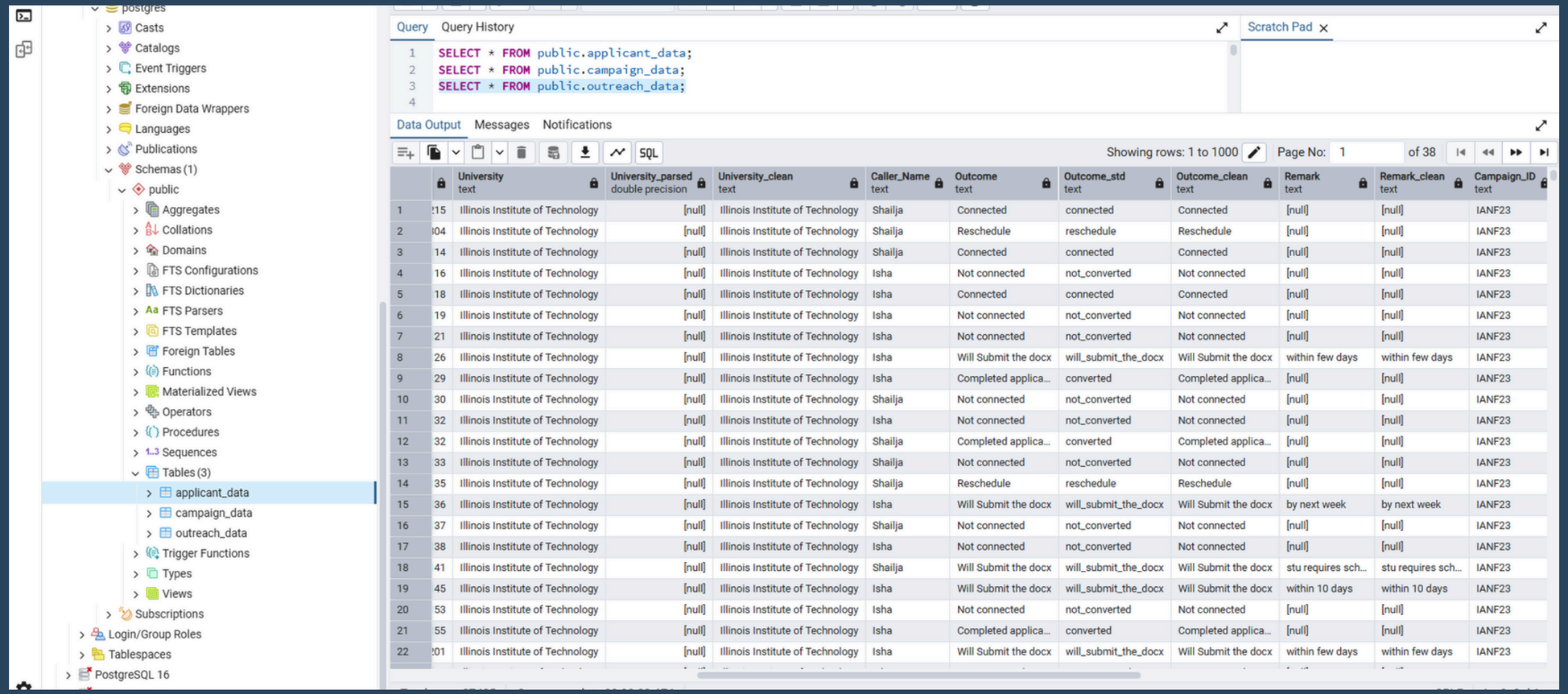
Campaign_Data



	ID text	ID_clean text	ID_num bigint	Name text	Name_parsed double precision	Category text	Category_clean text	Intake text	Intake_clean text	Intake_num bigint	University text
1	AANF23	AANF23	23	GR GS FA24 Campaign- Admit No Deposit	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
2	AND23	AND23	23	GR GS FA24 Campaign- Deposit No Action	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
3	BPNAN...	BPNANF...	23	GR GS FA24 Campaign- Deposit No I-20	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
4	BPNND...	BPNND23	23	GR GS FA24 Campaign- In Progress	[null]	Pre Admission	Pre Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
5	CTKANF...	CTKANF23	23	GR GS FA24 Campaign- Submit Incomplete	[null]	Pre Admission	Pre Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
6	DANE24	DANE24	24	GR GS Call Campaign: India ANF	[null]	Pre Admission	Pre Admission	AY2024	AY2024	2024	Illinois Institute of Technolog

Data Set Overview

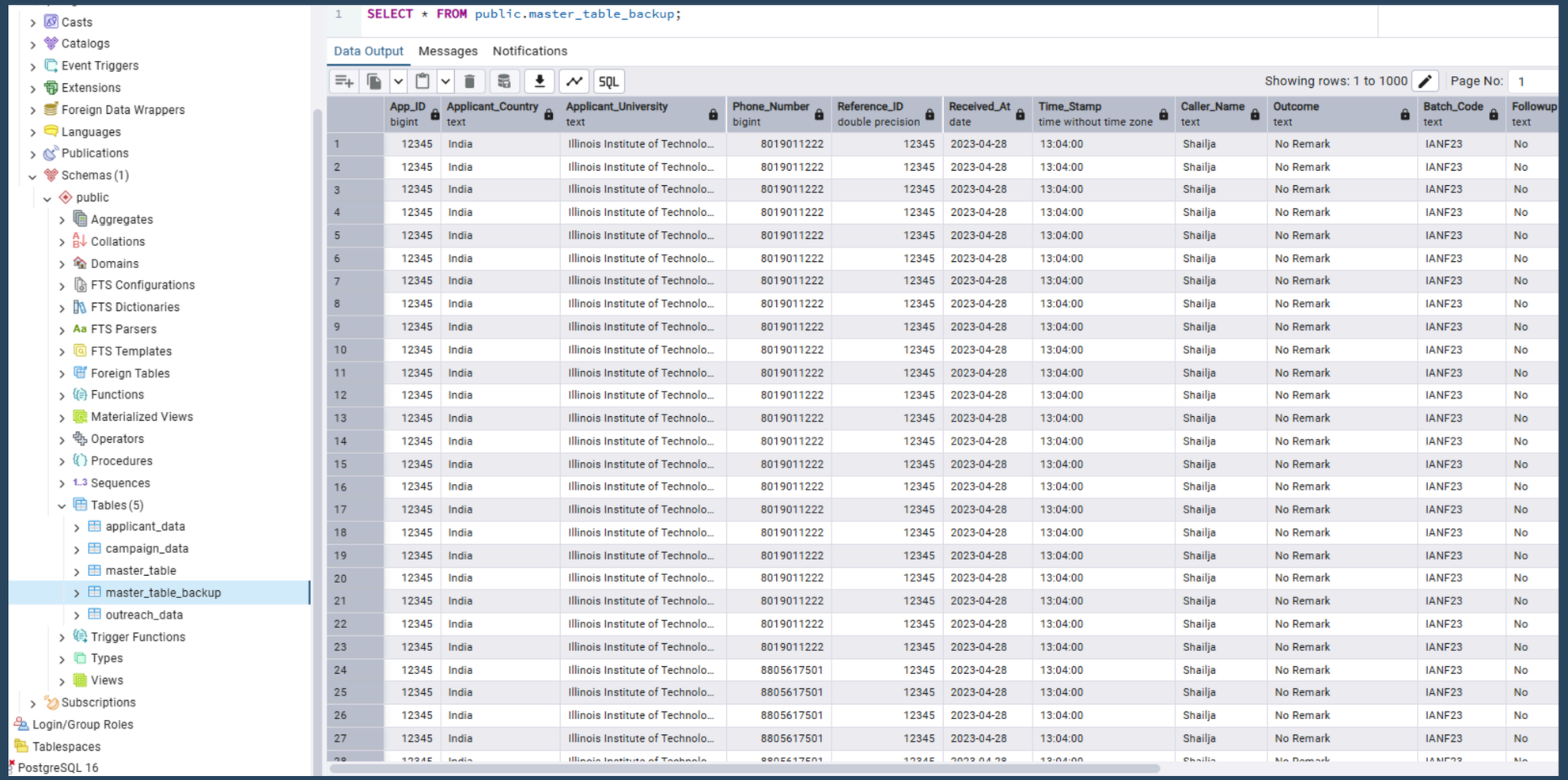
Outreach_Data



	University text	University_parsed double precision	University_clean text	Caller_Name text	Outcome text	Outcome_std text	Outcome_clean text	Remark text	Remark_clean text	Campaign_ID text
1	215	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Shailja	Connected	connected	Connected	[null]	[null]	IANF23
2	104	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Shailja	Reschedule	reschedule	Reschedule	[null]	[null]	IANF23
3	14	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Shailja	Connected	connected	Connected	[null]	[null]	IANF23
4	16	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Not connected	not_converted	Not connected	[null]	[null]	IANF23
5	18	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Connected	connected	Connected	[null]	[null]	IANF23
6	19	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Not connected	not_converted	Not connected	[null]	[null]	IANF23
7	21	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Not connected	not_converted	Not connected	[null]	[null]	IANF23
8	26	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Will Submit the docx	will_submit_the_docx	Will Submit the docx	within few days	within few days	IANF23
9	29	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Completed applica...	converted	Completed applica...	[null]	[null]	IANF23
10	30	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Shailja	Not connected	not_converted	Not connected	[null]	[null]	IANF23
11	32	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Not connected	not_converted	Not connected	[null]	[null]	IANF23
12	32	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Shailja	Completed applica...	converted	Completed applica...	[null]	[null]	IANF23
13	33	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Shailja	Not connected	not_converted	Not connected	[null]	[null]	IANF23
14	35	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Shailja	Reschedule	reschedule	Reschedule	[null]	[null]	IANF23
15	36	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Will Submit the docx	will_submit_the_docx	Will Submit the docx	by next week	by next week	IANF23
16	37	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Shailja	Not connected	not_converted	Not connected	[null]	[null]	IANF23
17	38	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Not connected	not_converted	Not connected	[null]	[null]	IANF23
18	41	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Shailja	Will Submit the docx	will_submit_the_docx	Will Submit the docx	stu requires sch...	stu requires sch...	IANF23
19	45	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Will Submit the docx	will_submit_the_docx	Will Submit the docx	within 10 days	within 10 days	IANF23
20	53	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Not connected	not_converted	Not connected	[null]	[null]	IANF23
21	55	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Completed applica...	converted	Completed applica...	[null]	[null]	IANF23
22	201	Illinois Institute of Technology	[null]	Illinois Institute of Technology	Isha	Will Submit the docx	will_submit_the_docx	Will Submit the docx	within few days	within few days	IANF23

Data Set Overview

Master_Table



	App_ID bigint	Applicant_Country text	Applicant_University text	Phone_Number bigint	Reference_ID double precision	Received_At date	Time_Stamp time without time zone	Caller_Name text	Outcome text	Batch_Code text	Followup text
1	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
2	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
3	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
4	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
5	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
6	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
7	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
8	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
9	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
10	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
11	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
12	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
13	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
14	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
15	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
16	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
17	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
18	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
19	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
20	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
21	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
22	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
23	12345	India	Illinois Institute of Technolo...	8019011222	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
24	12345	India	Illinois Institute of Technolo...	8805617501	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
25	12345	India	Illinois Institute of Technolo...	8805617501	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
26	12345	India	Illinois Institute of Technolo...	8805617501	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No
27	12345	India	Illinois Institute of Technolo...	8805617501	12345	2023-04-28	13:04:00	Shailja	No Remark	IANF23	No

Exploratory Data Analysis (EDA) Steps

Key initial checks and data cleaning techniques used

In this section, we detail **initial checks** performed during EDA, including row counts and schema inspections. We also cover the **cleaning processes** for App_ID, Country, Phone Numbers, Campaign, and Outreach Data to ensure data quality and consistency.

Exploratory Data Analysis (EDA) Process Overview

```
------ Cleaning Applicant Data--------------------


UPDATE "applicant_data"
SET "App_ID" = REGEXP_REPLACE("App_ID", '[^0-9]', '', 'g');

SELECT "App_ID"
FROM "applicant_data"
WHERE "App_ID" = '' OR "App_ID" IS NULL;

DELETE FROM "applicant_data"
WHERE "App_ID" IS NULL
   OR TRIM("App_ID") = '';

ALTER TABLE "applicant_data"
ALTER COLUMN "App_ID" TYPE BIGINT USING "App_ID"::BIGINT;


SELECT "Phone_Number"
FROM "applicant_data"
WHERE "applicant_data" = "applicant_data";

ALTER TABLE "applicant_data"
ALTER COLUMN "Phone_Number" TYPE BIGINT
USING "Phone_Number"::BIGINT;
```

Clean App_ID - Remove all unwanted characters (like letters or symbols) and keep only numbers in the App_ID column.

Find empty App_IDs - Check if there are any rows where App_ID is missing or blank.

Delete invalid rows - Remove rows from the table where App_ID is missing or blank.

Fix App_ID format - Change the App_ID column type to a number format (BIGINT) so it’s stored properly as numeric data.

Fix Phone_Number format - Change the Phone_Number column type to a number format (BIGINT) so only valid numeric phone numbers are stored.

Exploratory Data Analysis (EDA) Process Overview

```
---------Replace placeholders with NULL-------------------------
UPDATE "applicant_data" SET "App_ID" = NULL
WHERE lower(trim(coalesce("App_ID"::text, ''))) IN
('n/a','na','none','-','--','null','<na>','unknown','');


--------Convert scientific notation (e.g., 9.20E+11) into plain digits--
UPDATE "applicant_data"
SET "App_ID" = trim(to_char("App_ID"::numeric, 'FM99999999999999999999
WHERE "App_ID" ~ '^[0-9]+(\.[0-9]+)?[eE][+-]?[0-9]+$';

DELETE FROM "applicant_data"
WHERE "App_ID" IN ('16309553555', '97455626509');

-------------- Checking count of duplicate values in App_id column------

SELECT "App_ID" , COUNT(*) AS dup_count
FROM "applicant_data"
WHERE "App_ID"  IS NOT NULL
GROUP BY "App_ID"
HAVING COUNT(*) > 1
ORDER BY dup_count DESC;


DELETE FROM "applicant_data"
WHERE "App_ID" IS NULL
   OR trim("App_ID") = ''
   OR "App_ID" ~ '^[\.,]+$'    -- only dots or commas
   OR "App_ID" IN (
        '919000000000',
        '920000000000',
        '918000000000',
```

Replace placeholders with NULL - Change fake values like NA, None, -, Unknown in App_ID to NULL (treat them as missing).

Fix scientific notation - Convert IDs written in formats like 9.20E+11 into plain numbers (e.g., 920000000000)

Remove specific invalid IDs - Delete rows with known wrong App_IDs (like 16309553555, 97455626509).

Check duplicate App_IDs - Find IDs that appear more than once and show how many times they are repeated.

Remove invalid App_IDs - Delete rows where App_ID is missing, blank, just symbols (., ,,), or certain invalid numbers.

Exploratory Data Analysis (EDA) Process Overview

```
-------------------------UPDATION-------
UPDATE "applicant_data" SET "Country" = NULL
WHERE lower(trim(coalesce("Country"::text, ''))) IN
('n/a','na','none','-','--','null','<na>','unknown','');

UPDATE "applicant_data"
SET "Country" = trim(both '"' FROM trim("Country"))
WHERE "Country" IS NOT NULL;

UPDATE "applicant_data"
SET "Country" = 'Not Provided'
WHERE "Country" ~* '@';


UPDATE "applicant_data"
SET "Country" = 'Not Provided'
WHERE "Country" IS NULL;

--------- In Country field there are many Rows with Not Provided so we will drop those rows--------

DELETE FROM "applicant_data"
WHERE "Country" = 'Not Provided';

-----------------------------------------------------------------------------------------------

UPDATE "applicant_data" SET "Phone_Number" = NULL
WHERE lower(trim(coalesce("Phone_Number"::text, ''))) IN
('n/a','na','none','-','--','null','<na>','unknown','');

------- Normalize phone numbers (keep only digits)-------------------------
UPDATE "applicant_data"
SET "Phone_Number" = regexp_replace("Phone_Number", '\D', '', 'g')
WHERE "Phone_Number" IS NOT NULL;

------------keep last 10 digits if longer-------------------------
UPDATE "applicant_data"
SET "Phone_Number" = right("Phone_Number", 10)
WHERE length("Phone_Number") > 10;
```

Set Country to NULL - Replace fake values like NA, None, -, Unknown in the Country column with NULL.
Clean extra quotes - Remove unwanted double quotes (") around country names.

Mark invalid country entries - If a country field contains an email-like value (with @), set it as Not Provided.
Fill missing countries - Replace NULL values in the Country column with Not Provided.

Drop Not Provided countries - Delete all rows where Country is Not Provided since they don't give useful information.

Set Phone_Number to NULL - Replace fake values like NA, None, -, Unknown in the Phone_Number column with NULL.
Normalize phone numbers - Remove all non-digit characters (like spaces, dashes, or symbols), keeping only numbers.

Exploratory Data Analysis (EDA) Process Overview

```
------Count of phone number where values are less than 10---------------
SELECT COUNT(*) AS invalid_phone_count
FROM "applicant_data"
WHERE LENGTH(TRIM("Phone_Number")) < 10;

-- 776 Phone numbers are invalid format(Either we can drop those rows or put {invalid})---
-- We will make Phone number field as INT so we cant store string values, so its better to drop those field.--

DELETE FROM "applicant_data"
WHERE LENGTH(TRIM("Phone_Number")) < 10;

-------------------------Cleaning Campaign Data-----------------------------------------
-- uppercase, remove non-alphanumeric characters
UPDATE "campaign_data"
SET "ID" = UPPER(REGEXP_REPLACE("ID", '[^A-Za-z0-9]', '', 'g'))
WHERE "ID" IS NOT NULL;


SELECT "ID", COUNT(*)
FROM "campaign_data"
GROUP BY "ID"
ORDER BY COUNT(*) DESC;

UPDATE "campaign_data"
SET "Name" = NULLIF(REGEXP_REPLACE(TRIM("Name"), '\s+', ' ', 'g'), '')
WHERE "Name" IS NOT NULL;


UPDATE "campaign_data"
SET "Name" = regexp_replace(
    "Name",
    '( - Admit.*| - No Deposit.*| - Deferrals.*| - In Progress.*| - Enrolled.*
    |Deposit and Advised.*|Bangladesh|Pakistan|Nepal|Africa|China|Taiwan|Korea)',
    '',
    'gi'
);
```

Count invalid phone numbers - Find how many phone numbers have less than 10 digits (result = 776 invalid numbers).

Delete invalid phone numbers - Remove all rows where the phone number has fewer than 10 digits, since they are incomplete.

Clean Campaign IDs - Convert ID values to uppercase and remove any special characters or spaces.

Check duplicate Campaign IDs - Count how many times each campaign ID appears to identify duplicates.

Clean Campaign Names - Remove unnecessary extra spaces and set empty names to NULL.

Simplify Campaign Names - Remove extra details like Admit, No Deposit, Enrolled, Deferrals, country names etc., so the campaign name is clean and standardized.

Exploratory Data Analysis (EDA) Process Overview

```sql
ALTER TABLE "campaign_data"
ADD COLUMN "Stage" TEXT;



-------extracting stage -----------

UPDATE "campaign_data"
SET "Stage" = CASE
    WHEN "Name" ILIKE '%No Deposit%' THEN 'No Deposit'
    WHEN "Name" ILIKE '%Admit%' THEN 'Admit'
    WHEN "Name" ILIKE '%Deposit%' THEN 'Deposit'
    WHEN "Name" ILIKE '%Deferral%' THEN 'Deferral'
    WHEN "Name" ILIKE '%Enrolled%' THEN 'Enrolled'
    WHEN "Name" ILIKE '%In Progress%' THEN 'In Progress'
    ELSE 'Completed'
END;

UPDATE "campaign_data"
SET "Stage" = 'Incomplete'
WHERE "ID" = 'DANE24';

UPDATE "campaign_data"
SET "Stage" = 'In Progress'
WHERE "ID" = 'CTKANF23';

UPDATE "campaign_data"
SET "Stage" = 'Incomplete'
WHERE "ID" = 'OND23';

UPDATE "campaign_data"
SET "Stage" = 'New Inquiry'
WHERE "ID" = 'OANF23';


UPDATE "campaign_data"
```

Add Stage column - Add a new column called “Stage” to store the campaign status separately from the campaign name.

Extract Stage from Name - Look at the campaign name, find keywords like “No Deposit,” “Admit,” “Deposit,” “Deferral,” “Enrolled,” or “In Progress,” and fill the Stage column accordingly (anything else is marked as “Completed”).

Update Stage for DANE24 - Set the stage as “Incomplete” for the campaign with ID DANE24.
Update Stage for CTKANF23 - Set the stage as “In Progress” for the campaign with ID CTKANF23.
Update Stage for OND23 - Set the stage as “Incomplete” for the campaign with ID OND23.

Update Stage for OANF23 - Set the stage as “New Inquiry” for the campaign with ID OANF23.
Update Stage for SP25NIQ - Set the stage as “Not Enrolled” for the campaign with ID SP25NIQ.

Exploratory Data Analysis (EDA) Process Overview

```sql
CREATE TABLE "master_table" AS
SELECT
    a."App_ID",
    a."Country" AS "Applicant_Country",
    a."University" AS "Applicant_University",
    a."Phone_Number",

    o."Reference_ID",
    o."Received_At",
    o."Time_Stamp",
    o."Caller_Name",
    o."Outcome",
    o."Batch_Code",
    o."Followup",
    o."Remark",
    o."University" AS "Outreach_University",

    c."ID" AS "Campaign_ID",
    c."Name" AS "Campaign_Name",
    c."Category" AS "Campaign_Category",
    c."Intake" AS "Campaign_Intake",
    c."University" AS "Campaign_University",
    c."Status" AS "Campaign_Status",
    c."Start_Date",
    c."Stage"
FROM "applicant_data" a
LEFT JOIN "outreach_data" o
    ON a."App_ID" = o."Reference_ID"::BIGINT
LEFT JOIN "campaign_data" c
    ON o."University" = c."University";

SELECT * FROM "master_table";
```

Create master_table - Combine data from applicants, outreach, and campaigns into a single table called master_table.

Select Applicant Info - Take applicant ID, country, university, and phone number from applicant_data.

Select Outreach Info - Take reference ID, received date, timestamp, caller name, outcome, batch code, follow-up, remark, and university from outreach_data.

Select Campaign Info - Take campaign ID, name, category, intake, university, status, start date, and stage from campaign_data.

Join Applicant & Outreach - Link applicants to outreach records where the applicant ID matches the outreach reference ID.

Join Applicant & Outreach - Link applicants to outreach records where the applicant ID matches the outreach reference ID.

Data Validation Process Overview

Key steps ensuring data integrity and accuracy

The data validation process includes crucial steps: handling duplicates, treating missing values, and applying business rule checks. These actions ensure that the dataset remains reliable, consistent, and ready for analysis, ultimately supporting informed decision-making.

Validation Process Overview

```
-- New Metrics
CASE
    WHEN o."Outcome" IN ('Interested','Not Interested','Pending Decision',
                         'Visa Related','Application In Progress',
                         'Pending Documents','Scholarship Related',
                         'Loan/Bank Document Query','I-20 Related Issue')
    THEN 1 ELSE 0 END AS "Connected_Calls",

CASE
    WHEN o."Outcome" IN ('No Response','Disconnected','Invalid Number')
    THEN 1 ELSE 0 END AS "Disconnected_Calls",

CASE
    WHEN (CASE WHEN o."Outcome" IN ('Interested','Not Interested','Pending Decision',
                                    'Visa Related','Application In Progress',
                                    'Pending Documents','Scholarship Related',
                                    'Loan/Bank Document Query','I-20 Related Issue')
               THEN 1 ELSE 0 END
          +
          CASE WHEN o."Outcome" IN ('No Response','Disconnected','Invalid Number')
               THEN 1 ELSE 0 END) > 0
    THEN ROUND(
        (CASE WHEN o."Outcome" IN ('Interested','Not Interested','Pending Decision',
                                   'Visa Related','Application In Progress',
                                   'Pending Documents','Scholarship Related',
                                   'Loan/Bank Document Query','I-20 Related Issue')
              THEN 1 ELSE 0 END)::numeric
        /
        ((CASE WHEN o."Outcome" IN ('Interested','Not Interested','Pending Decision',
                                    'Visa Related','Application In Progress',
                                    'Pending Documents','Scholarship Related',
                                    'Loan/Bank Document Query','I-20 Related Issue')
               THEN 1 ELSE 0 END)
         +
         (CASE WHEN o."Outcome" IN ('No Response','Disconnected','Invalid Number')
               THEN 1 ELSE 0 END)),2)
    ELSE NULL END AS "Connectivity_Rate",

CASE
    WHEN (CASE WHEN o."Outcome" IN ('Interested','Not Interested','Pending Decision',
                                    'Visa Related','Application In Progress',
                                    'Pending Documents','Scholarship Related',
                                    'Loan/Bank Document Query','I-20 Related Issue')
```

```
               +
               CASE WHEN o."Outcome" IN ('No Response','Disconnected','Invalid Number')
                    THEN 1 ELSE 0 END) > 0
          THEN CASE
              WHEN ROUND((CASE WHEN o."Outcome" IN ('Interested','Not Interested','Pending Decision',
                                                    'Visa Related','Application In Progress',
                                                    'Pending Documents','Scholarship Related',
                                                    'Loan/Bank Document Query','I-20 Related Issue')
                               THEN 1 ELSE 0 END)::numeric
                         /
                         ((CASE WHEN o."Outcome" IN ('Interested','Not Interested','Pending Decision',
                                                     'Visa Related','Application In Progress',
                                                     'Pending Documents','Scholarship Related',
                                                     'Loan/Bank Document Query','I-20 Related Issue')
                                THEN 1 ELSE 0 END)
                          +
                          (CASE WHEN o."Outcome" IN ('No Response','Disconnected','Invalid Number')
                                THEN 1 ELSE 0 END)),2) >= 0.7
                  THEN 'Good'
              WHEN ROUND((CASE WHEN o."Outcome" IN ('Interested','Not Interested','Pending Decision',
                                                    'Visa Related','Application In Progress',
                                                    'Pending Documents','Scholarship Related',
                                                    'Loan/Bank Document Query','I-20 Related Issue')
                               THEN 1 ELSE 0 END)::numeric
                         /
                         ((CASE WHEN o."Outcome" IN ('Interested','Not Interested','Pending Decision',
                                                     'Visa Related','Application In Progress',
                                                     'Pending Documents','Scholarship Related',
                                                     'Loan/Bank Document Query','I-20 Related Issue')
                                THEN 1 ELSE 0 END)
                          +
                          (CASE WHEN o."Outcome" IN ('No Response','Disconnected','Invalid Number')
                                THEN 1 ELSE 0 END)),2) >= 0.4
                  THEN 'Average'
              ELSE 'Poor'
          END
    ELSE NULL END AS "Agent_Performance"

FROM "applicant_data" a
LEFT JOIN "outreach_data" o
    ON a."App_ID" = o."Reference_ID"::BIGINT
LEFT JOIN "campaign_data" c
    ON o."University" = c."University";
```

Validation Process Overview

Created Some New Metrics

Connected_Calls - Count calls where the outcome shows any engagement like Interested, Pending Decision, Visa Related, Application In Progress, etc.

- Shows how many applicants were successfully reached by the agent.
- Helps understand which calls generated a meaningful response.

Connectivity_Rate → Calculate the ratio of connected calls to total calls (connected + disconnected) and round it to 2 decimal places.

- Measures how effectively agents are reaching applicants.
- Higher rate = more successful communication with applicants.

Disconnected_Calls - Count calls where there was no response, the number was invalid, or the call got disconnected.

- Helps identify failed outreach attempts.
- Useful for spotting issues with phone numbers or call quality.

Agent_Performance - Categorize the agent based on connectivity rate:

- 'Good' if ≥ 70% of calls were connected,
- 'Average' if 40–69% were connected,
- 'Poor' if < 40%,
- NULL if there were no calls.

Validation Process Overview

```
----- Deleting those fields where the app id is more than 6 in length as it is inappropriate---

DELETE FROM "master_table"
WHERE LENGTH("App_ID"::text) <> 6;

------- Dropped some inappropriate column which is not in use for any analysis case---

ALTER TABLE "master_table"
DROP COLUMN "Phone_Number",
DROP COLUMN "Reference_ID",
DROP COLUMN "Batch_Code",
DROP COLUMN "Followup",
DROP COLUMN "Outreach_University",
DROP COLUMN "Campaign_ID",
DROP COLUMN "Campaign_Status",
DROP COLUMN "Campaign_University";

------ creating some new metrics like connected, disconnected calls, and
-- agent performance as average poor, etc.-----

UPDATE "master_table"
SET "Agent_Performance" = CASE
    WHEN ("Connected_Calls" + "Disconnected_Calls") > 0
    THEN CASE
        WHEN ROUND("Connected_Calls"::numeric / NULLIF(("Connected_Calls" + "Disconnected_Calls"),0), 2) >= 0.7
            THEN 'Good'
        WHEN ROUND("Connected_Calls"::numeric / NULLIF(("Connected_Calls" + "Disconnected_Calls"),0), 2) >= 0.4
            THEN 'Average'
        ELSE 'Poor'
    END
    ELSE 'Not Evaluated'    -- fallback when no calls
END;


----- deleteing field where values are null. this can sause problems in analysis ----

DELETE FROM "master_table"
WHERE "Received_At" IS NULL
```

Delete invalid App_IDs - Remove rows where the applicant ID is not exactly 6 characters long, as these are considered invalid.

Drop unused columns - Remove columns like Phone_Number, Reference_ID, Batch_Code, Followup, Outreach_University, Campaign_ID, Campaign_Status, and Campaign_University because they are not needed for analysis.

Update Agent_Performance → Categorize agents based on call success:

- 'Good' if ≥ 70% of calls were connected,
- 'Average' if 40–69% connected,
- 'Poor' if < 40%,
- 'Not Evaluated' if no calls exist for that agent.

Delete rows with null values → Remove rows where Received_At or Time_Stamp is NULL, as missing data can cause problems in analysis.

Validation Process Overview

```sql
-- Performing validation

-- Count total rows in cleaned table

SELECT COUNT(*) AS total_records
FROM "master_table";



-- Check duplicate rows based on unique identifier

SELECT "App_ID", COUNT(*)
FROM "master_table"
GROUP BY "App_ID"
HAVING COUNT(*) > 1;



-- Remove duplicate rows from master_table_cleaned

DELETE FROM "master_table"
WHERE ctid NOT IN (
    SELECT ctid
    FROM (
        SELECT ctid,
               ROW_NUMBER() OVER (
                   PARTITION BY "App_ID"
                   ORDER BY "Time_Stamp" DESC   -- keep latest row
               ) AS rn
        FROM "master_table"
    ) t
    WHERE rn = 1
);


-- Count missing values column-wise
SELECT
    SUM(CASE WHEN "App_ID" IS NULL THEN 1 ELSE 0 END) AS missing_app_id,
    SUM(CASE WHEN "Applicant_Country" IS NULL THEN 1 ELSE 0 END) AS missing_country,
    SUM(CASE WHEN "Applicant_University" IS NULL THEN 1 ELSE 0 END) AS missing_university,
    SUM(CASE WHEN "Outcome" IS NULL THEN 1 ELSE 0 END) AS missing_outcome,
    SUM(CASE WHEN "Campaign_Name" IS NULL THEN 1 ELSE 0 END) AS missing_campaign
FROM "master_table";
```

Count total records - Find the total number of rows in the cleaned master table to see how many records are available for analysis.

Check duplicate App_IDs - Identify applicants that appear more than once by counting rows for each App_ID.

Remove duplicate rows - Keep only the latest record per App_ID based on Time_Stamp and delete older duplicates to ensure data accuracy.

Count missing values - Check how many rows have missing values in key columns like App_ID, Applicant_Country, Applicant_University, Outcome, and Campaign_Name to identify gaps in the data.

Validation Process Overview

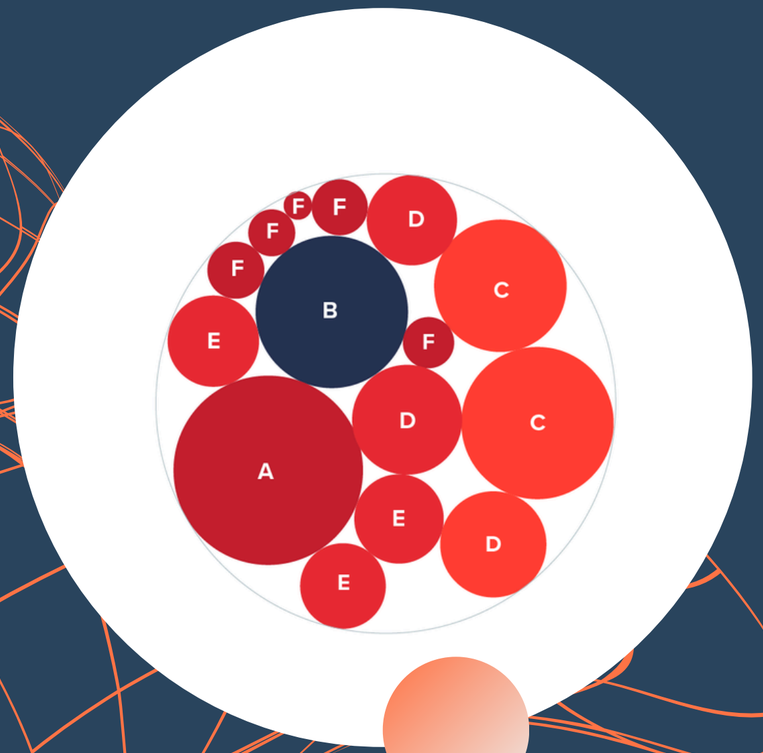
```sql
----Checking Connectivity Rate is between 0 and 1
SELECT COUNT(*) AS invalid_connectivity
FROM "master_table"
WHERE "Connectivity_Rate" < 0 OR "Connectivity_Rate" > 1;


--- Checking Agen performance label

SELECT DISTINCT "Agent_Performance"
FROM "master_table";


-- Checking for unexpected Stage values

SELECT DISTINCT "Stage"
FROM "master_table";


-- List unique call outcomes

SELECT DISTINCT "Outcome"
FROM "master_table";
```

Check Connectivity_Rate - Count how many rows have a connectivity rate less than 0 or greater than 1 to identify any invalid calculations.

Check Agent_Performance labels - List all unique values in the Agent_Performance column to verify that labels like Good, Average, Poor, or Not Evaluated are correct.

Check Stage values - List all unique values in the Stage column to ensure only expected stages like No Deposit, Admit, Deposit, etc., are present.

Check call outcomes - List all unique values in the Outcome column to see all types of call results recorded and identify any unexpected or incorrect entries.



Final Master Data Overview

Consolidated and validated for analysis readiness

The **Final Master Data** integrates all relevant datasets, ensuring they are **clean, validated,** and consistent. This step is crucial for accurate visualization, KPI measurement, and downstream analytics, enabling data-driven decision-making for stakeholders.

Final Master Data After Validation Overview

Data Output | Messages | Notifications

Showing rows: 1 to 594 | Page No: 1 of 1

	App_ID bigint	Applicant_Country text	Applicant_University text	Received_At date	Time_Stamp time without time zone	Caller_Name text	Outcome text	Remark text	Campaign_Name text	Campaign_Category text	Can text
1	427192	India	Illinois Institute of Technology	2024-03-20	18:22:00	Rudra	No Response	not connected	GR GS SP25 Campaign	Pre Admission	AY2
2	471044	Ghana	Illinois Institute of Technology	2024-11-05	18:51:00	Rudra	No Response	student will join sp25 session	GR GS FA24 Campaign	Post Admission	AY2
3	429501	India	Illinois Institute of Technology	2024-03-28	18:34:00	Rudra	Pending Documents	will submit the docx	GR GS FA24 Campaign	Pre Admission	AY2
4	450950	India	Illinois Institute of Technology	2024-06-05	12:47:00	Rudra	Interested	not interested to iit	GR GS Call Campaign	Post Admission	AY2
5	365548	India	Illinois Institute of Technology	2023-05-12	12:47:00	Poppy	Application In Progress	not interested	GR GS Call Campaign	Pre Admission	AY2
6	367708	India	Illinois Institute of Technology	2023-05-03	13:51:00	Namrata	I-20 Related Issue	completed application	Not Provided	Post Admission	AY2
7	375465	India	Illinois Institute of Technology	2023-05-11	16:16:00	Shailja	No Response	reschedule	GR GS FA24 Campaign	Post Admission	AY2
8	349472	India	Illinois Institute of Technology	2023-05-12	12:40:00	Poppy	Pending Documents	completed application	GR GS Call Campaign	Pre Admission	AY2
9	366971	India	Illinois Institute of Technology	2023-05-12	13:05:00	Shailja	I-20 Related Issue	completed application	GR GS FA24 Campaign	Post Admission	AY2
10	346519	India	Illinois Institute of Technology	2023-05-12	17:08:00	Poppy	Application In Progress	not interested	GR GS SP25 Campaign	Pre Admission	AY2
11	366637	India	Illinois Institute of Technology	2023-05-12	17:33:00	Poppy	Pending Documents	will submit the docx	GR GS SP25 Campaign	Pre Admission	AY2
12	355606	South Africa	Illinois Institute of Technology	2023-05-25	18:08:00	Poppy	Pending Documents	will submit the docx	GR GS FA24 Campaign	Post Admission	AY2
13	365504	India	Illinois Institute of Technology	2023-05-29	14:54:00	Shailja	I-20 Related Issue	completed application	GR GS SP25 Campaign	Pre Admission	AY2
14	355548	India	Illinois Institute of Technology	2023-07-03	16:54:00	Shailja	Visa Related	ready to pay the deposit	GR GS Call Campaign	Post Admission	AY2
15	351960	Pakistan	Illinois Institute of Technology	2023-07-05	17:49:00	Isha	Application In Progress	ready to pay the deposit	Not Provided	Post Admission	AY2
16	350602	Pakistan	Illinois Institute of Technology	2023-07-07	18:20:00	Poppy	Visa Related	want to defer	GR GS SP25 Campaign	Post Admission	AY2
17	450195	India	Illinois Institute of Technology	2024-03-18	15:50:00	Prajwal	Pending Documents	completed application	GR GS Call Campaign	Post Admission	AY2
18	358188	India	Illinois Institute of Technology	2023-05-01	13:56:00	Shailja	Interested	not interested	GR GS Call Campaign	Post Admission	AY2
19	434967	India	Illinois Institute of Technology	2024-03-21	14:13:00	Rudra	Pending Decision	will confirm later	Not Provided	Post Admission	AY2
20	410407	India	Illinois Institute of Technology	2024-03-21	14:15:00	Rudra	Pending Decision	will confirm later	GR GS SP25 Campaign	Pre Admission	AY2
21	427742	India	Illinois Institute of Technology	2024-03-21	14:16:00	Rudra	Pending Decision	will confirm later	GR GS FA24 Campaign	Pre Admission	AY2
22	423283	India	Illinois Institute of Technology	2024-03-21	14:28:00	Rudra	Pending Decision	will confirm later	Not Provided	Post Admission	AY2
23	447745	India	Illinois Institute of Technology	2024-03-26	13:19:00	Prajwal	Pending Documents	completed application	GR GS FA24 Campaign	Post Admission	AY2
24	450208	India	Illinois Institute of Technology	2024-03-29	18:08:00	Prajwal	Pending Documents	completed application	GR GS SP25 Campaign	Post Admission	AY2
25	452094	India	Illinois Institute of Technology	2024-03-29	18:28:00	Rudra	Pending Documents	will submit the docx	GR GS Call Campaign	Pre Admission	AY2
26	450333	India	Illinois Institute of Technology	2024-03-29	19:01:00	Rudra	Pending Documents	will submit the docx	GR GS Call Campaign	Post Admission	AY2

Total rows: 594 | Query complete 00:00:02.865 | CRLF | Ln 1, Col 1

Before And After Validation Review on Master Data

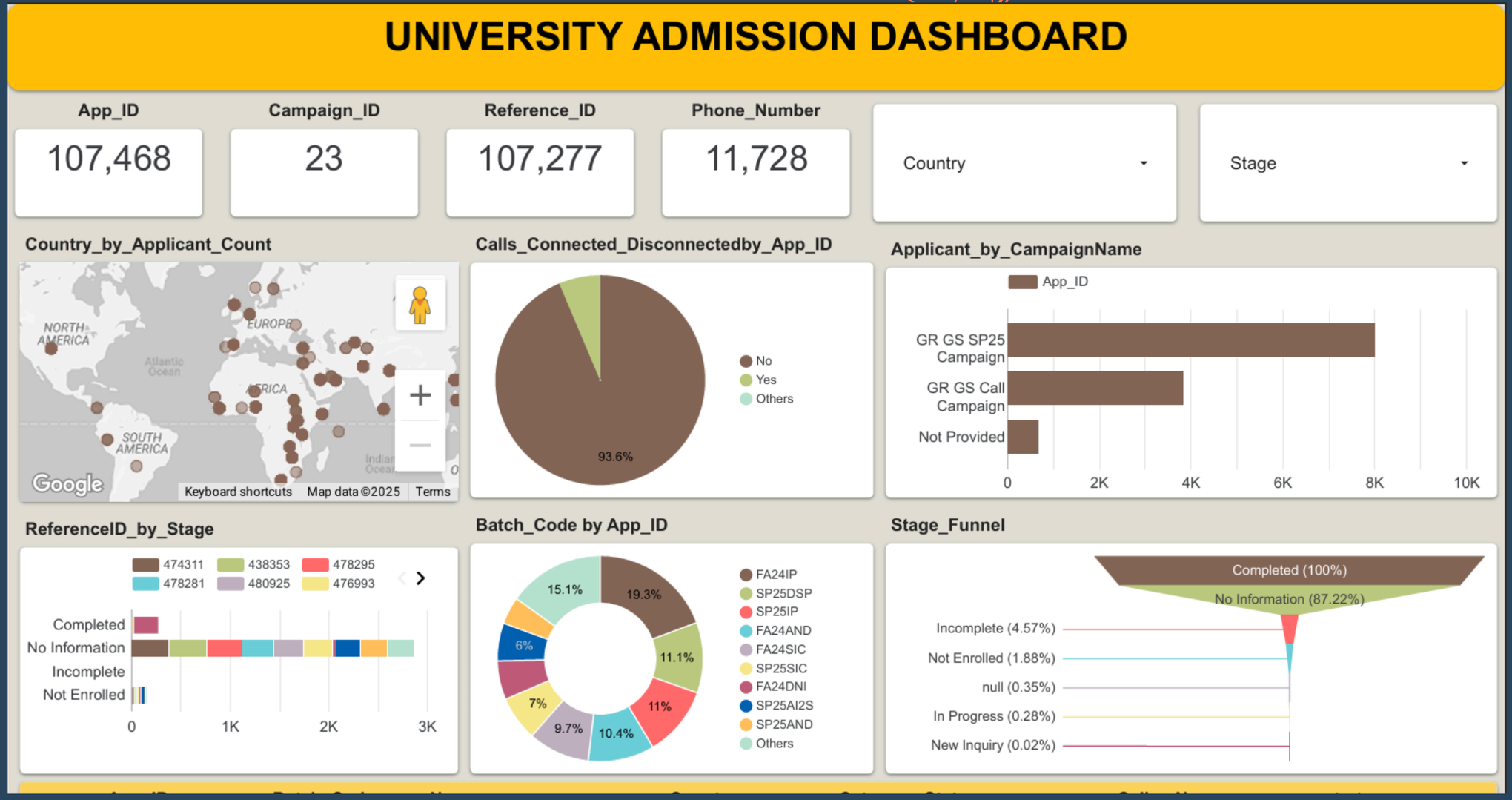
Before Validation

- **Size & Structure**
  - ~24 lakh rows (raw dataset).
  - Combined data from Applicant, Outreach, and Campaign tables.
  - Contained duplicates, invalid App_IDs, and columns with excessive null values.
- **Data Quality Issues**
  - Multiple records per applicant (e.g., Applicant ID 432571 had 69 rows).
  - Inflated outcomes like Application in Progress (11,178) and Pending Document (7,475).
  - High-null fields (Phone_Number, Reference_ID, Batch_Code, Campaign_Status, etc.).
  - Missing timestamps in several records.
  - Business rules not enforced (e.g., invalid App_ID formats, incorrect connectivity rates).

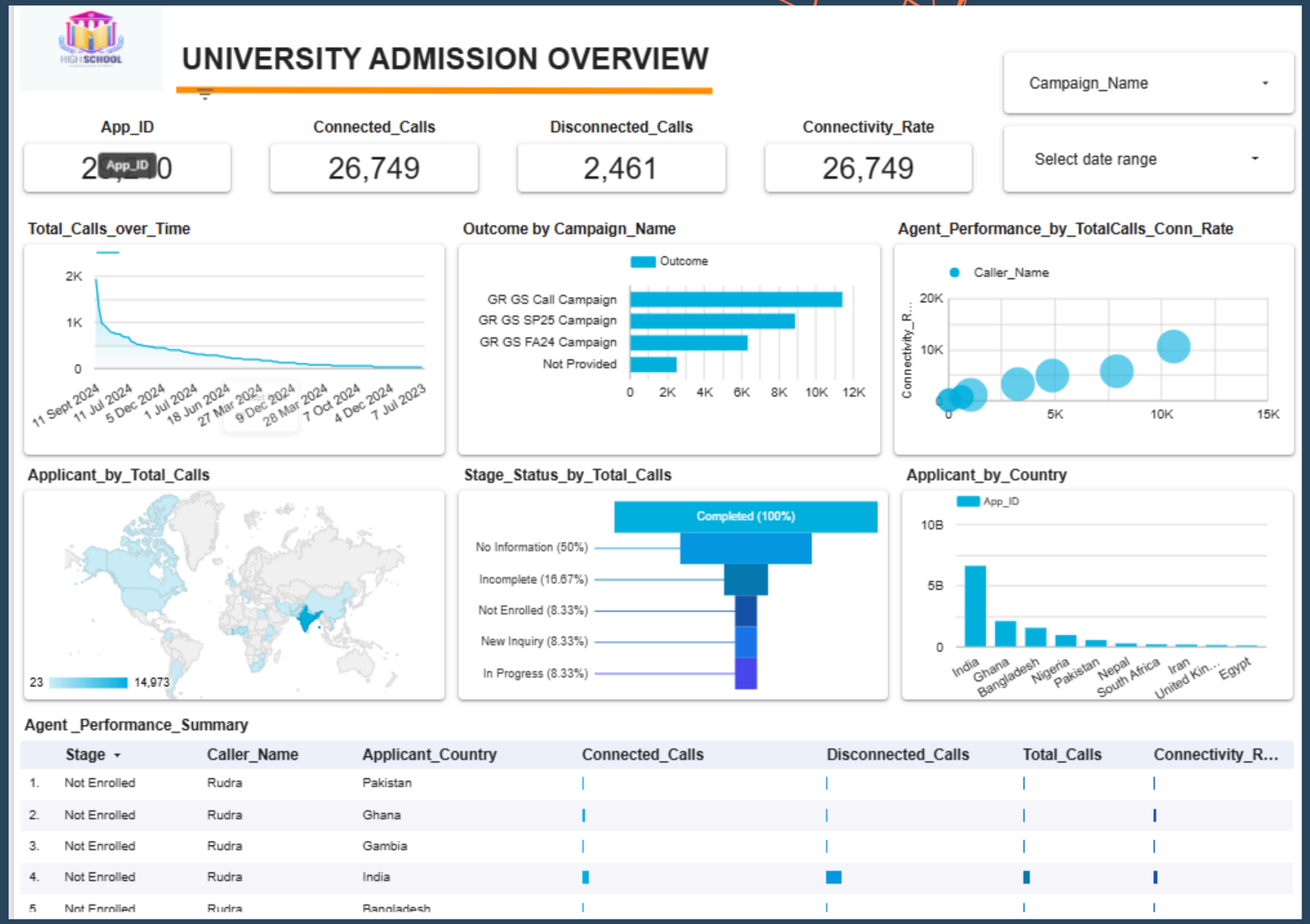
After Validation

- **Size & Structure**
  - Dataset reduced significantly after cleaning.
  - Only valid records retained with unique App_IDs.
  - High-null and redundant columns dropped.
- **Data Quality Improvements**
  - Each applicant has only 1 record (e.g., Applicant ID 432571 reduced from 69 → 1).
  - Correct outcome counts: Application in Progress (209) and Pending Document (192).
  - Removed rows with missing/invalid timestamps.
  - **Business rules applied:**
    - App_ID must be 6-digit.
    - Connectivity rates between 0 and 1.
    - Valid agent performance categories only (Good/Average/Poor/Not Evaluated).
    - Campaign values cross-checked with allowed categories.

Visualization Overview Before Validation



Visualization Overview After Validation



Major Differences Betwwen Before And After Validation Dashboards

Aspect	Before Validation	After Validation
Dataset Size	~24 lakh rows (raw, with duplicates, invalid IDs, high-null columns).	Reduced dataset, cleaned of duplicates & invalid IDs, only valid records kept.
Applicant Outcomes	“Application In Progress” = 11,178; “Pending Document” = 7,475 (inflated counts).	“Application In Progress” = 209; “Pending Document” = 192 (true values).
Applicant ID Level	Same applicant had multiple rows (e.g., ID 432571 had 69 records).	Each applicant now has only 1 valid record (ID 432571 → 1 record).
Calls Connected Data	Connectivity metrics overstated due to duplicate activity.	Duplicates removed, actual call data preserved (no data loss).
Stage Funnel	Funnel stages inflated (Incomplete, Pending, In Progress larger than actual).	Funnel stages accurate, reflecting true applicant distribution.
Campaign Analysis	Applicant counts exaggerated across campaigns.	Campaign visuals show correct applicant reach.
Batch Code Analysis	Batch allocations inflated by duplicate entries.	Clean batch distribution with authentic applicant counts.
Geographical Mapping	Applicant counts per country exaggerated due to duplicates.	Accurate applicant density per region.
Data Quality	High-null columns (Phone_Number, Reference_ID, etc.) created noise in visuals.	High-null columns dropped, visuals are cleaner and more reliable.

Key Recommendations for Data Quality

Focus on improving data accuracy and reliability in the organization.

Automate validation pipelines

Ensure consistent data checks and reduce manual intervention.

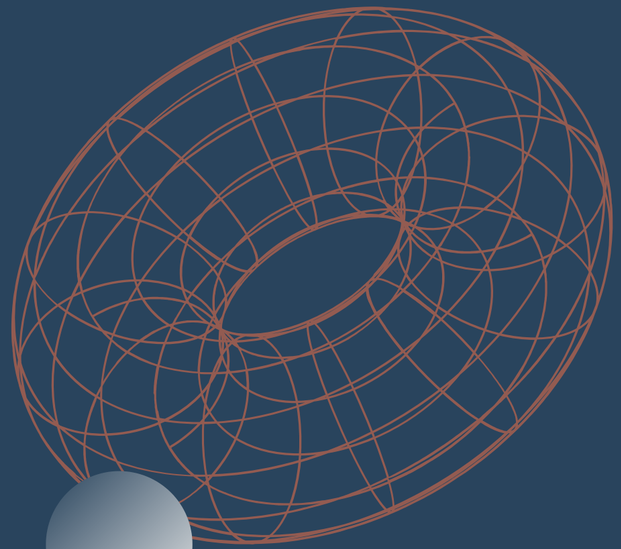
Establish data quality KPIs

Measure performance and maintain high standards of data integrity.

Enforce strong data entry controls

Prevent errors at the source to enhance overall data quality.

Importance of a Clean Dataset



A clean, reliable, and validated dataset is crucial for **supporting actionable insights** and informed decision-making. It ensures that analyses are based on accurate data, promoting efficiency and effectiveness in organizational strategies.

Any questions or thoughts?

We welcome your input and insights! Please share.