Team Charter - TEAM 20

Team Members	Saurav Kumar (davidsauravyadav@gmail.com) Lingolu Alekhya (alekhyalingolu@gmail.com) Anjali Bista (bistaanjali1415@gmail.com) Praveen Kumar M (mpraveenkumar5397@gmail.com) Sharath Raju Saraswathil (sharathraju1230@gmail.com)
Team Lead	Saurav Kumar Email: davidsauravyadav@gmail.com
Team Member Roles and Responsibilities	**Sponsors: SLU** **Client: Excelerate** **Saurav Kumar** (davidsauravyadav@gmail.com) - **Project Lead:** Responsible for project planning, coordination, and overall delivery. **Anjali Bista** (bistaanjali1415@gmail.com) - **Decision Maker:** Ensures timely decision-making, sets goals, and reviews progress. **Praveen Kumar M** (mpraveenkumar5397@gmail.com) **- Communication Facilitator:** Manages communication channels, schedules meetings, and keeps stakeholders informed.
Mission, Vision, Objective & Core Values	**Mission:** To achieve strategic project goals efficiently and collaboratively while ensuring high standards and timely outcomes. **Vision:** To be recognised for outstanding teamwork, reliability, and innovation in every assigned project.

	Objectives: Deliver all key milestones within defined timelines, foster a culture of accountability and positive collaboration, and continually seek improvement. **Core Values:** Integrity, Accountability, Discipline, Respect, Innovation, Collaboration.
Internal Checks, Milestones, and Reviews	Regular progress monitoring, periodic review meetings, and clear milestone tracking to maintain momentum and identify bottlenecks early. Emphasis on structured feedback and documentation for continuous improvement and knowledge sharing. Proactive risk management and openness to feedback.
Operations	**Meetings:** Weekly status review and planning sessions. TCM, Weekly meetups and Group calls. **Assignments:** Clear allocation of responsibilities and deadlines. **Documentation:** All records accessible centrally; revision logs maintained. **Status Updates:** Routine updates provided for transparency. **Deadlines:** Week 1, deliverable to be posted on or before 11:59 pm on 8th Sept. and week 1 started from 1st Sept.
Continuous Learning & Development	To foster innovation and adaptability, the team commits to ongoing skill enhancement through regular workshops, online courses, and peer knowledge-sharing sessions. Team members are encouraged to pursue individual learning goals that align with project needs, including emerging technologies, leadership, and domain expertise. Reflection meetings to be held twice a week to know the project's growth and to solve issues among team members within the project.

NAME : SAURAV KUMAR

ROLE: DATA VISUALIZATION ASSOCIATE INTERNSHIP

1. Objective

The primary objective of this project was to perform an exploratory data analysis (EDA) and comprehensive data cleaning on three distinct datasets: ApplicantData, CampaignData, and OutreachData, and perform some visualization. The cleaned and preprocessed data was then to be loaded into a PostgreSQL database for further query-based analysis.

2. Data Used

The following three datasets were used for this analysis:

- **ApplicantData.csv:** Contains information on applicants, including App_ID, Country, University, and Phone_Number. It has 37,882 entries and one missing value in the App_ID column.

- **CampaignData.csv:** Contains data about various campaigns with columns such as ID, Name, Category, Intake, University, Status, and Start_Date. This dataset has 23 entries with no missing values.

- **OutreachData.csv:** Contains details of outreach efforts, including Reference_ID, Recieved_At_University, Caller_Name, Outcome, Remark, Campaign_ID, and Escalation_Required. It has 37,881 entries and a significant number of missing values in the Remark column.

```
ApplicantData.csv - Head:
   App_ID  Country                    University  Phone_Number
0   12345    India  Illinois Institute of Technology    9823241234
1   12345    India  Illinois Institute of Technology    8805617501
2   12345    India  Illinois Institute of Technology   18019011222
3  347397  Nigeria  Illinois Institute of Technology    7738599513
4  347397  Nigeria  Illinois Institute of Technology  919182706838

ApplicantData.csv - Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37882 entries, 0 to 37881
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   App_ID        37881 non-null  object
 1   Country       37882 non-null  object
 2   University    37882 non-null  object
 3   Phone_Number  37882 non-null  object
dtypes: object(4)
memory usage: 1.2+ MB
None

CampaignData.csv - Head:
        ID                                    Name        Category  Intake  \
0    AANF23   GR GS FA24 Campaign- Admit, No Deposit  Post Admission  AY2024
1     AND23   GR GS FA24 Campaign- Deposit No Action  Post Admission  AY2024
2  BPNANF23    GR GS FA24 Campaign- Deposit, No I-20  Post Admission  AY2024
3   BPNND23         GR GS FA24 Campaign- In Progress   Pre Admission  AY2024
4  CTKANF23  GR GS FA24 Campaign- Submit, Incomplete   Pre Admission  AY2024

                     University     Status       Start_Date
0  Illinois Institute of Technology  Completed   3/20/2024 0:00
1  Illinois Institute of Technology  Completed  9/11/2024 00:00
2  Illinois Institute of Technology  Completed  7/11/2024 00:00
3  Illinois Institute of Technology  Completed   3/6/2024 00:00
4  Illinois Institute of Technology  Completed   3/8/2024 00:00

CampaignData.csv - Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23 entries, 0 to 22
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   ID          23 non-null     object
 1   Name        23 non-null     object
 2   Category    23 non-null     object
 3   Intake      23 non-null     object
 4   University  23 non-null     object
 5   Status      23 non-null     object
 6   Start_Date  23 non-null     object
dtypes: object(7)
memory usage: 1.4+ KB
None

OutreachData.csv - Head:
  Reference_ID     Recieved_At                    University Caller_Name  \
0        12345  4/28/2023 12:15  Illinois Institute of Technology     Shailja
1        12345  4/28/2023 13:04  Illinois Institute of Technology     Shailja
2        12345   5/1/2023 11:14  Illinois Institute of Technology     Shailja
3       347397   5/1/2023 11:16  Illinois Institute of Technology        Isha
4       347397   5/1/2023 11:18  Illinois Institute of Technology        Isha

        Outcome Remark Campaign_ID Escalation_Required
0     Connected    NaN      IANF23                  No
1    Reschedule    NaN      IANF23                  No
2     Connected    NaN      IANF23                  No
3  Not connected    NaN      IANF23                  No
4     Connected    NaN      IANF23                  No

OutreachData.csv - Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37881 entries, 0 to 37880
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
```

3. Data Cleaning and Preprocessing

The data cleaning and preprocessing pipeline involved several key steps to ensure data quality and consistency:

- **Loading and Initial Inspection**: The three datasets were loaded into Pandas DataFrames. An initial inspection was performed to check the shape, data types, and identify missing values.
- **Helper Functions**: A suite of helper functions was created to automate common cleaning tasks. These included:
    - clean_phone(): Extracted the last 10 digits from the Phone_Number column, returning NaN for invalid entries.
    - remove_emails_and_clean_name(): Removed email addresses and performed basic cleanup on name-like columns by stripping whitespace and removing commas.
    - standardize_outcome(): Mapped various outcome strings to a consistent set of labels like converted, not_converted, and pending.
- **Automated Cleaning Pipeline**: A clean_dataframe function was created to apply these cleaning steps systematically across all datasets. It performed the following actions:
    - **Column Renaming**: Replaced spaces in column names with underscores.
    - **Whitespace and Missing Values**: Trimmed whitespace from object columns and replaced empty strings with NaN.
    - **Data Type Conversion**: Parsed date columns and converted numeric-like columns from object to numeric types.
    - **Duplicate Removal**: Dropped exact duplicate rows from the dataframes.
    - **Outlier Handling**: Applied IQR capping to handle outliers in numeric columns.
- **Result of Cleaning**: The cleaning process successfully reduced the ApplicantData from 37,882 rows to 19,997 rows by removing duplicates.
- **EDA and Cleaning Pipeline:** The EDA and cleaning pipeline was an automated and modular process designed to prepare the three datasets (**ApplicantData, CampaignData, and OutreachData**) for analysis. The pipeline involved several key steps:

  Helper Functions: You created reusable functions to handle specific cleaning tasks. For example, clean_phone() standardized phone numbers, remove_emails_and_clean_name() cleaned text fields, and standardize_outcome() mapped different string values to a consistent set of categories.

  Automated Cleaning: A main clean_dataframe function was developed to apply a series of cleaning steps systematically. This function performed column name standardization, handled whitespace, converted data types, and removed duplicate rows.

  Duplicate Removal: A crucial step was the identification and removal of duplicate rows. This was particularly effective on the ApplicantData, where it reduced the number of rows from 37,882 to 19,997, ensuring a dataset without redundant entries.

```
4      Connected    NaN      JANF23                 No

OutreachData.csv - Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37881 entries, 0 to 37880
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Reference_ID        37881 non-null  object
 1   Recieved_At         37881 non-null  object
 2   University          37881 non-null  object
 3   Caller_Name         37881 non-null  object
 4   Outcome             37881 non-null  object
 5   Remark              4077 non-null   object
 6   Campaign_ID         37881 non-null  object
 7   Escalation_Required 37881 non-null  object
dtypes: object(8)
memory usage: 2.3+ MB
None
```

```
Loaded C:\\Users\\decent\\OneDrive\\Desktop\\Excelerate_Project_Week1\\ApplicantData.csv with shape (37882, 4)

Running EDA...

=== EDA for ApplicantData ===
Shape: (37882, 4)

Dtypes:
 App_ID          object
Country          object
University       object
Phone_Number     object
dtype: object

Missing (by column):
App_ID          1
Country         0
University      0
```

```
Sample values for some object columns:
  - App_ID: ['12345', '12345', '12345', '347397', '347397']
  - Country: ['India', 'India', 'India', 'Nigeria', 'Nigeria']
  - University: ['Illinois Institute of Technology', 'Illinois Institute of Technology', 'Illinois Institute of Technology', 'Illinois Institute of T
chnology', 'Illinois Institute of Technology']
  - Phone_Number: ['9823241234', '8805617501', '18019011222', '7738599513', '919182706838']

Numeric summary (safe):
No numeric summary available or no numeric columns.

C:\Users\decent\AppData\Local\Temp\ipykernel_44616\3138405261.py:50: UserWarning: The argument 'infer_datetime_format' is deprecated and will be remo
ed in a future version. A strict version of it is now the default, see https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html. You
an safely remove this argument.
  df[c + '_parsed'] = pd.to_datetime(df[c], errors='coerce', infer_datetime_format=True)
Saved cleaned file to: C:\Users\decent\OneDrive\Desktop\Excelerate_Project_Week1\cleaned_ApplicantData.csv

Cleaning summary for ApplicantData:
Original shape: (37882, 4)
```

```
Cleaning summary for ApplicantData:
Original shape: (37882, 4)
After drop duplicates: (19997, 13)
Phone columns detected: ['Phone_Number']
Date columns parsed: ['University']
Name columns cleaned: []
Outcome/status columns standardized: []

Sample cleaned rows for ApplicantData:
```

	App_ID	App_ID_clean	App_ID_num	Country	Country_clean	University	University_parsed	University_clean	Phone_Number	Phone_Number_clean	Phone_Num
0	12345	12345	309963.0	India	India	Illinois Institute of Technology	NaT	Illinois Institute of Technology	9823241234	9823241234	9.82
1	12345	12345	309963.0	India	India	Illinois Institute of Technology	NaT	Illinois Institute of Technology	8805617501	8805617501	8.80

```
Loaded C:\\Users\\decent\\OneDrive\\Desktop\\Excelerate_Project_Week1\\CampaignData.csv with shape (23, 7)

Running EDA...

=== EDA for CampaignData ===
Shape: (23, 7)

Dtypes:
 ID            object
Name           object
Category       object
Intake         object
University     object
Status         object
Start_Date     object
dtype: object
```

```
 Missing (by column):
 ID           0
 Name         0
 Category     0
 Intake       0
 University   0
 Status       0
 Start_Date   0
 dtype: int64

 Sample values for some object columns:
   - ID: ['AANF23', 'AND23', 'BPNANF23', 'BPNND23', 'CTKANF23']
   - Name: ['GR GS FA24 Campaign- Admit, No Deposit', 'GR GS FA24 Campaign- Deposit No Action', 'GR GS FA24 Campaign- Deposit, No I-20', 'GR GS FA24
mpaign- In Progress', 'GR GS FA24 Campaign- Submit, Incomplete']
   - Category: ['Post Admission', 'Post Admission', 'Post Admission', 'Pre Admission', 'Pre Admission']
   - Intake: ['AY2024', 'AY2024', 'AY2024', 'AY2024', 'AY2024']
   - University: ['Illinois Institute of Technology', 'Illinois Institute of Technology', 'Illinois Institute of Technology', 'Illinois Institute of
chnology', 'Illinois Institute of Technology']
```

```
chnology', 'Illinois Institute of Technology']
  - Status: ['Completed', 'Completed', 'Completed', 'Completed', 'Completed']
  - Start_Date: ['3/20/2024 0:00', '9/11/2024 00:00', '7/11/2024 00:00', '3/6/2024 00:00', '3/8/2024 00:00']

Numeric summary (safe):
No numeric summary available or no numeric columns.
Saved cleaned file to: C:\Users\decent\OneDrive\Desktop\Excelerate_Project_Week1\cleaned_CampaignData.csv

Cleaning summary for CampaignData:
Original shape: (23, 7)
After drop duplicates: (23, 24)
Phone columns detected: []
Date columns parsed: ['Start_Date', 'Name', 'University']
Name columns cleaned: ['Name', 'Name_parsed']
Outcome/status columns standardized: ['Status', 'Status_clean']
```

	ID	ID_clean	ID_num	Name	Name_parsed	Category	Category_clean	Intake	Intake_clean	Intake_num	...	Status_std	Status_clean	Start_Date
0	AANF23	AANF23	23	GR GS FA24 Campaign- Admit No Deposit	NaT	Post Admission	Post Admission	AY2024	AY2024	2024	...	converted	Completed	3/20/2024 0:00
1	AND23	AND23	23	GR GS FA24 Campaign- Deposit No Action	NaT	Post Admission	Post Admission	AY2024	AY2024	2024	...	converted	Completed	9/11/2024 00:00
				GR GS FA24										

```
Loaded C:\\Users\\decent\\OneDrive\\Desktop\\Excelerate_Project_Week1\\OutreachData.csv with shape (37881, 8)

Running EDA...

=== EDA for OutreachData ===
Shape: (37881, 8)

Dtypes:
 Reference_ID          object
Recieved_At            object
University             object
Caller_Name            object
Outcome                object
Remark                 object
Campaign_ID            object
Escalation_Required    object
dtype: object

Missing (by column):
```

```
Missing (by column):
Remark               33804
Reference_ID             0
Recieved_At              0
University               0
Caller_Name              0
Outcome                  0
Campaign_ID              0
Escalation_Required      0
dtype: int64

Sample values for some object columns:
  - Reference_ID: ['12345', '12345', '12345', '347397', '347397']
  - Recieved_At: ['4/28/2023 12:15', '4/28/2023 13:04', '5/1/2023 11:14', '5/1/2023 11:16', '5/1/2023 11:18']
  - University: ['Illinois Institute of Technology', 'Illinois Institute of Technology', 'Illinois Institute of Technology', 'Illinois Institute o
chnology', 'Illinois Institute of Technology']
  - Caller_Name: ['Shailja', 'Shailja', 'Shailja', 'Isha', 'Isha']
  - Outcome: ['Connected', 'Reschedule', 'Connected', 'Not connected', 'Connected']
```

```
Saved cleaned file to: C:\Users\decent\OneDrive\Desktop\Excelerate_Project_Week1\cleaned_OutreachData.csv

Cleaning summary for OutreachData:
Original shape: (37881, 8)
After drop duplicates: (37435, 25)
Phone columns detected: []
Date columns parsed: ['Recieved_At', 'University']
Name columns cleaned: ['Caller_Name']
Outcome/status columns standardized: ['Outcome', 'Outcome_clean']

Sample cleaned rows for OutreachData:
```

	Reference_ID	Reference_ID_clean	Reference_ID_num	Recieved_At	Recieved_At_parsed	Recieved_At_clean	Recieved_At_num	University	University_parsed	Univer
0	12345	12345	348763.5	4/28/2023 12:15	2023-04-28 12:15:00	4/28/2023 12:15	42820231215	Illinois Institute of Technology	NaT	Illinoi of T

```
=== Final cleaned datasets summary ===
ApplicantData: shape=(19997, 13)
Top missing columns:
 University_parsed         19997
App_ID_num                 2715
App_ID_clean_num           2715
Phone_Number               1054
Phone_Number_num           1054
Phone_Number_clean_num     1054
App_ID                        1
App_ID_clean                  0
dtype: int64
----------------------------------------
CampaignData: shape=(23, 24)
Top missing columns:
 Name_parsed          23
University_parsed     23
ID                     0
Status                 0
Intake_clean_num       0
ID_clean_num           0
Status_clean_std       0
Start_Date_num         0
dtype: int64
----------------------------------------
OutreachData: shape=(37435, 25)
Top missing columns:
 University_parsed            37435
Remark                        33358
Reference_ID_num               3879
Reference_ID_clean_num         3879
Reference_ID                      0
Recieved_At_clean_num             0
Outcome_clean_std                 0
Escalation_Required_clean         0
dtype: int64
----------------------------------------

Common column names across datasets (lowercased): {'university', 'university_parsed', 'university_clean'}
No obvious ID key found for automatic merge. You can merge manually with a specified key.

Done. Cleaned files written to /mnt/data/ (filenames starting with 'cleaned_').
```

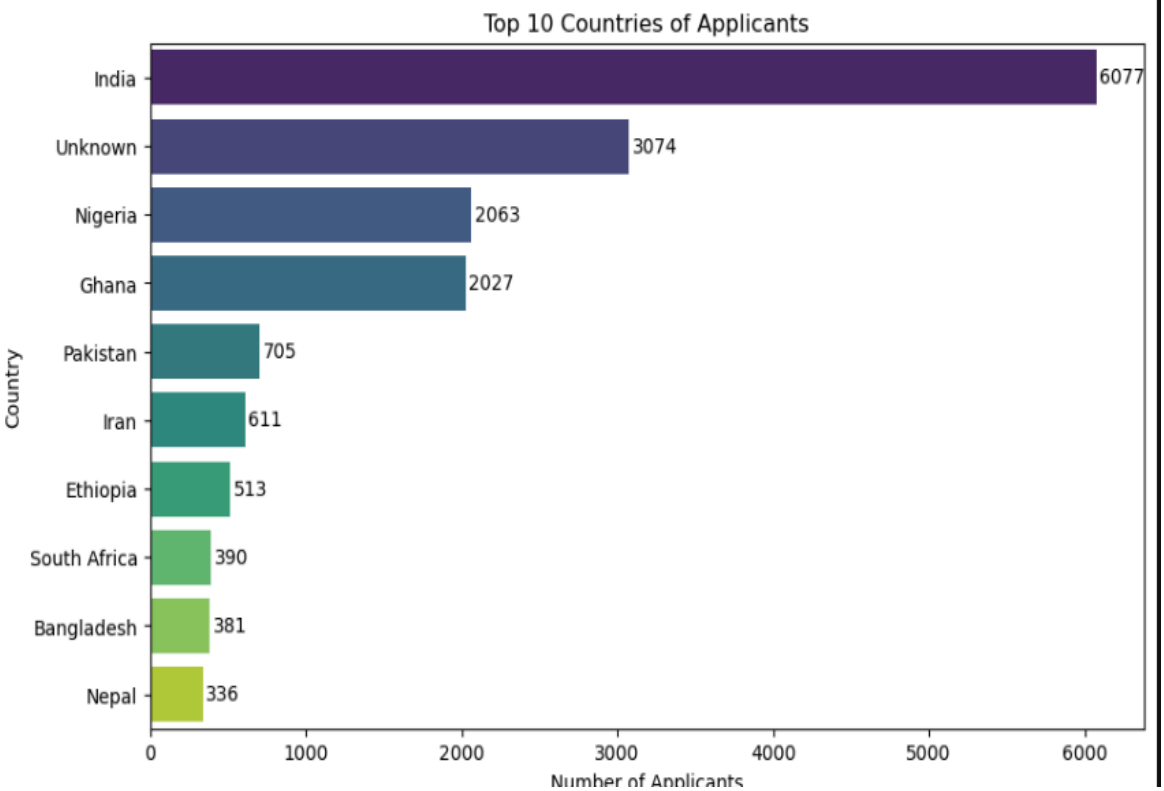
4. Data Visualization

The following visualizations provide a detailed overview of the datasets, revealing key insights into applicant demographics, campaign performance, and outreach efforts.

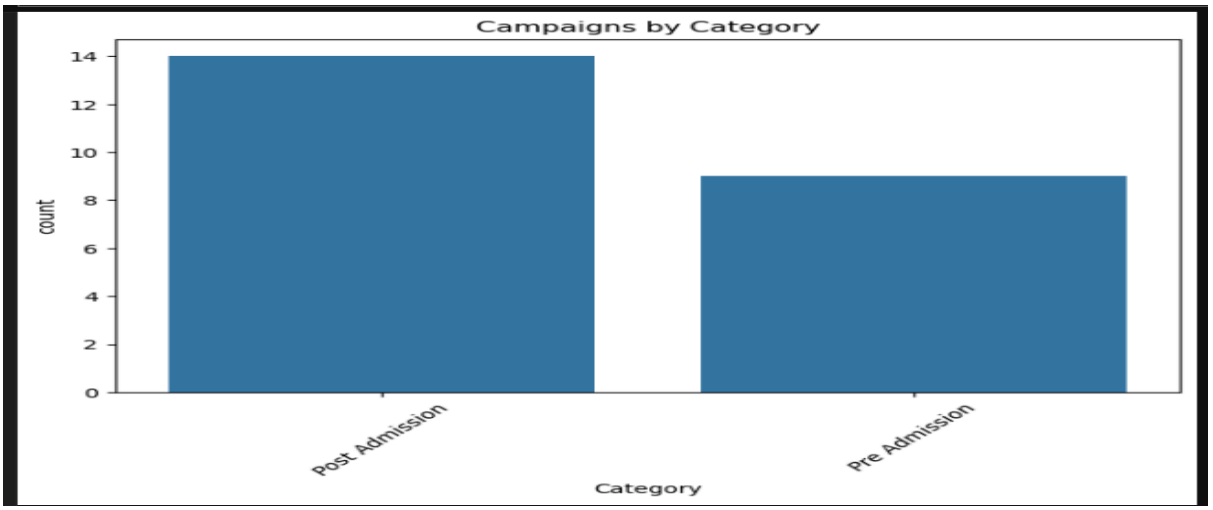
- **Applicants by Country**: This chart shows the top 10 countries with the most applicants, with **India** having the highest count.

```
       Country  Count
0        India   6077
1      Unknown   3074
2      Nigeria   2063
3        Ghana   2027
4     Pakistan    705
5         Iran    611
6     Ethiopia    513
7 South Africa    390
8   Bangladesh    381
9        Nepal    336
```

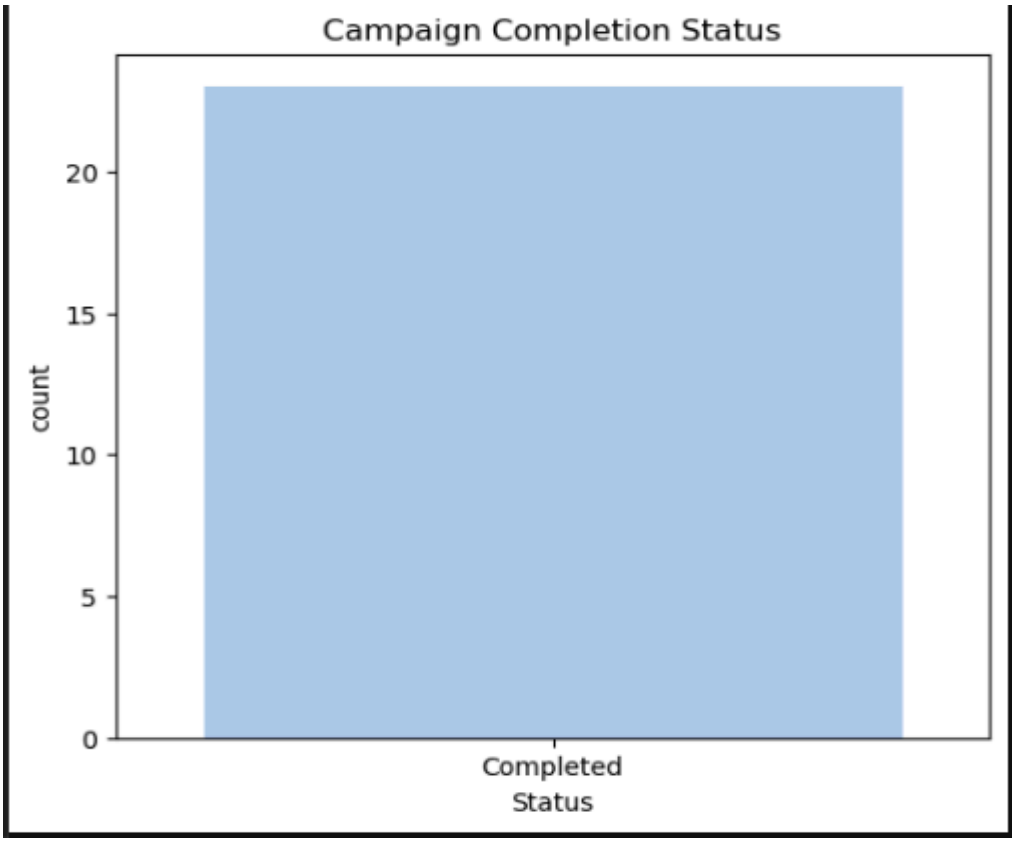
```
C:\Users\decent\AppData\Local\Temp\ipykernel_44616\2361198695.py:24: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and se
same effect.

  sns.barplot(data=country_counts, x="Count", y="Country", palette="viridis")
```



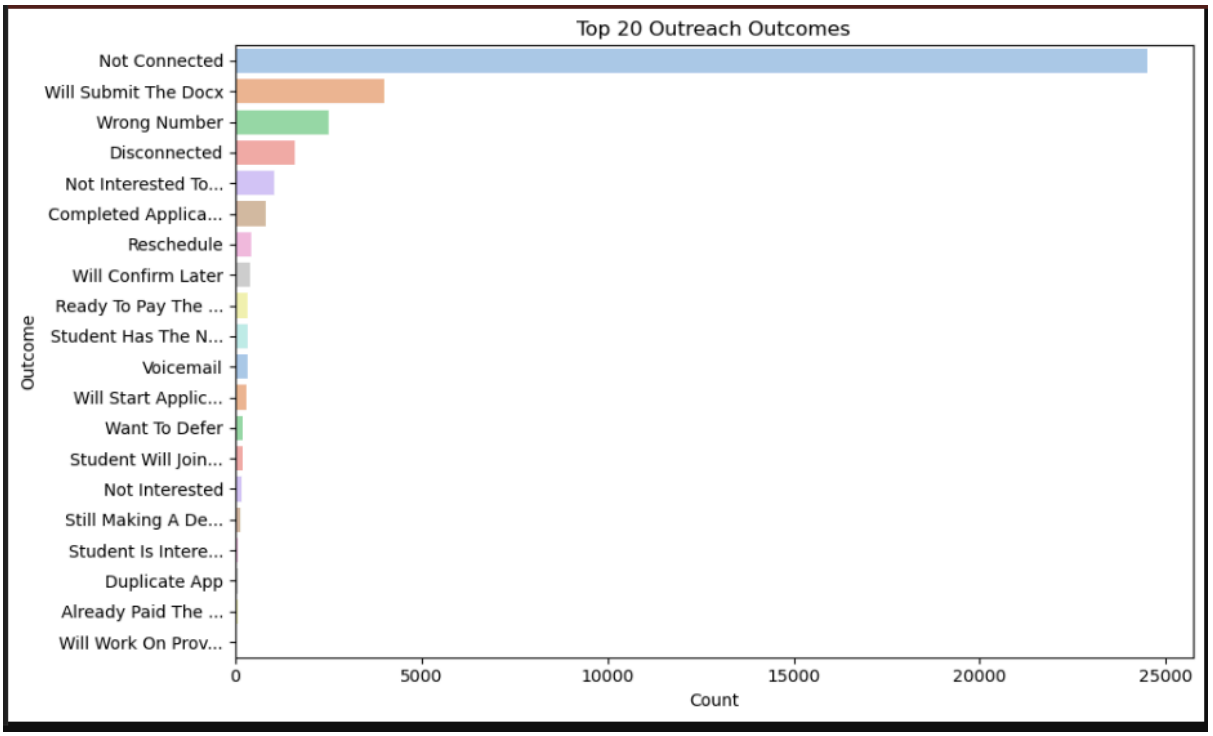
- **Campaigns by Category**: This visualization breaks down campaigns by category, revealing the most common types of campaigns.



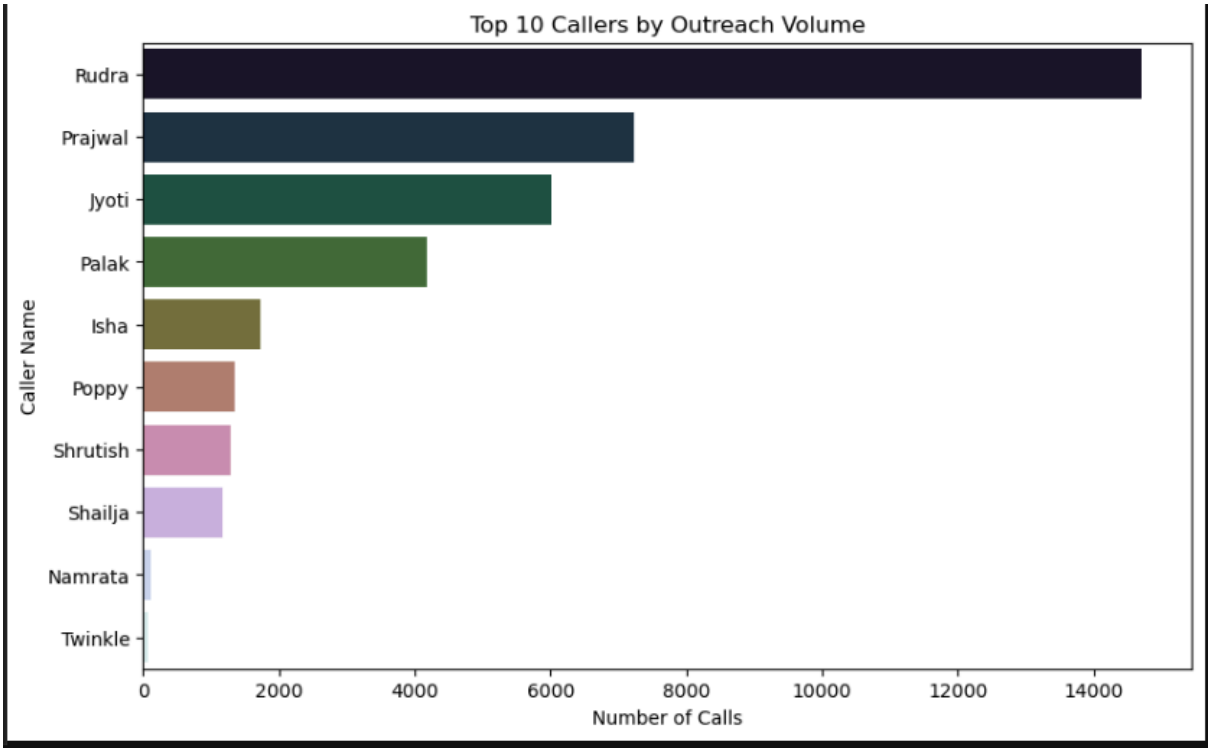
- **Campaign Status**: This chart displays the Campaign Completed Status



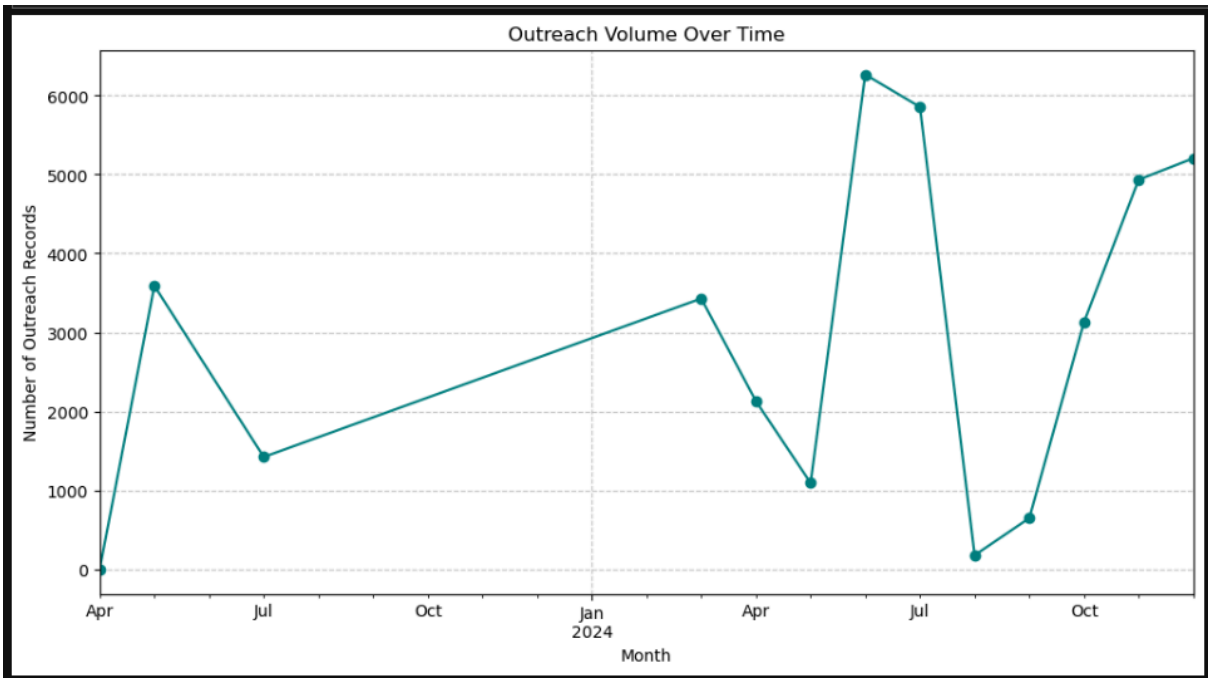
- **Outreach Outcomes**: This bar chart highlights the top 20 results from outreach efforts, such as whether the applicant was **Converted** or **Not Converted**.



- **Outreach by Caller**: This visualization ranks the top 10 callers by the number of outreach calls made, identifying the most active members of your team.



- **Outreach Trend Over Time**: This line plot illustrates the volume of outreach activity month by month, showing any **spikes** or **dips** in activity over time.

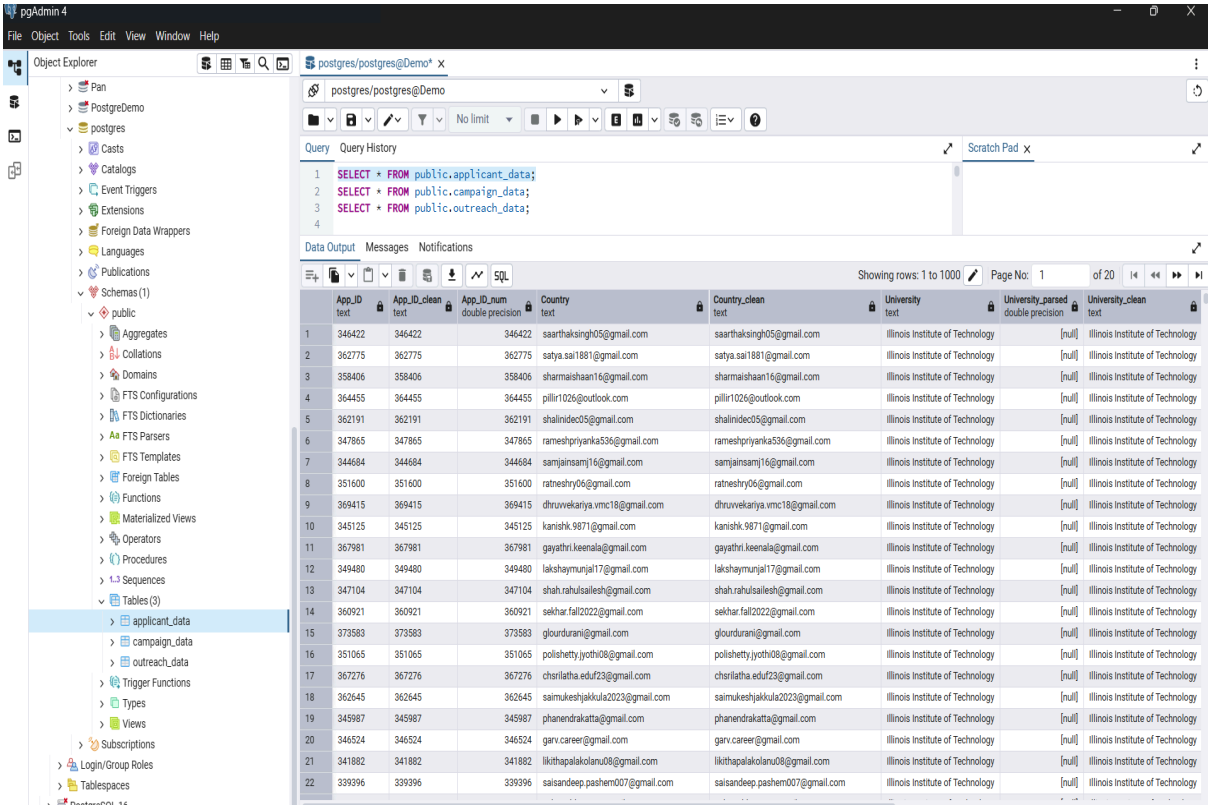


---

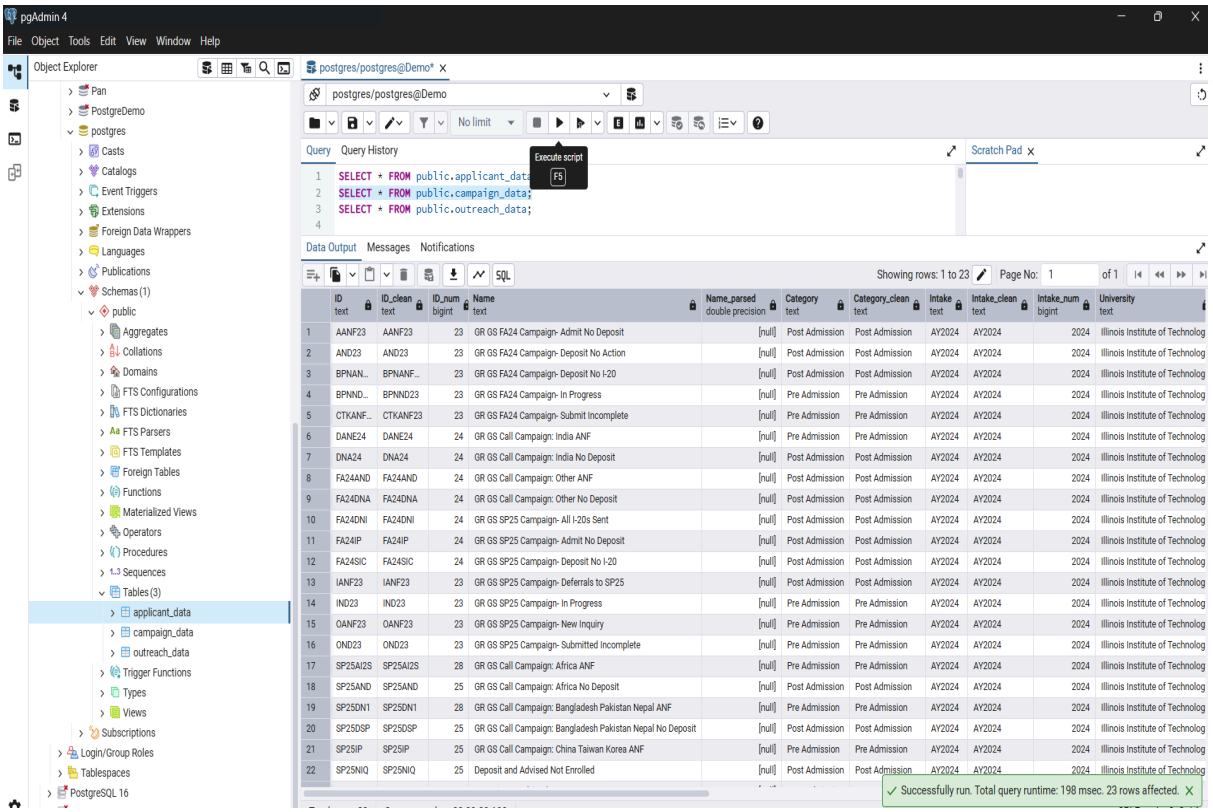
5. Loaded Cleaned Data in PostgreSQL using Python Script

A Python script was used to load the cleaned data into a PostgreSQL database. The script established a connection to the database and then used the df.to_sql() method to load each cleaned dataframe into its own table. The script successfully loaded the three datasets into the PostgreSQL database, replacing any existing tables.

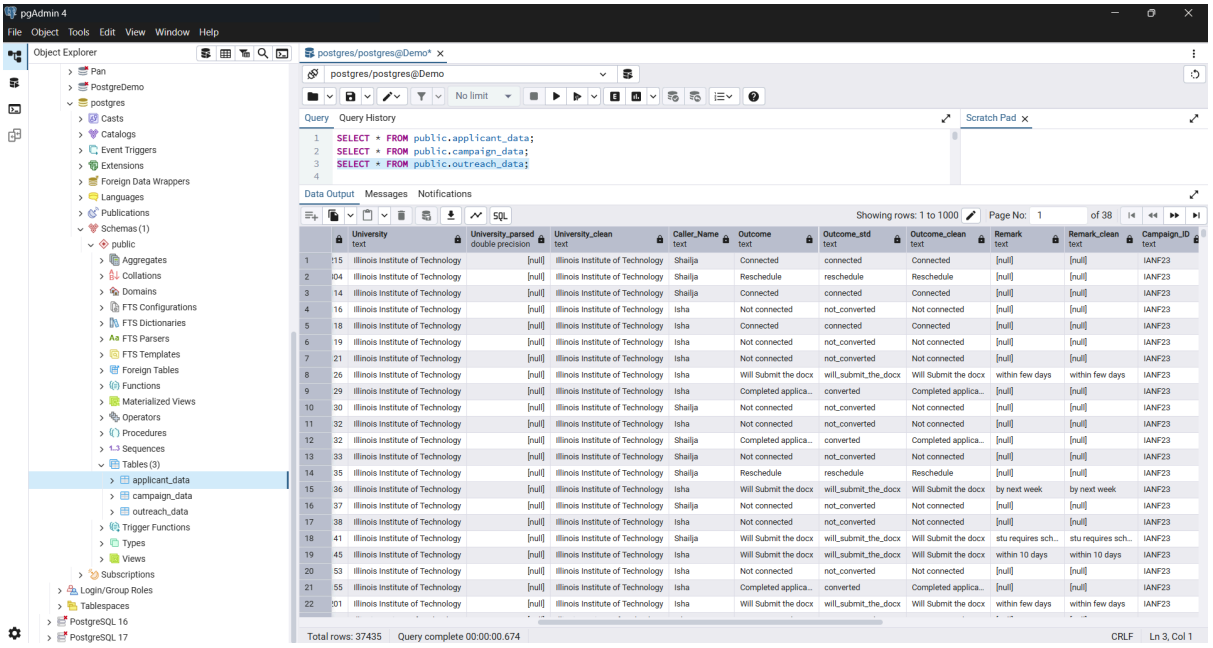
```
Loading C:\\Users\\decent\\OneDrive\\Desktop\\Excelerate_Project_Week1\\cleaned_ApplicantData.csv into table 'applicant_data'...
applicant_data loaded successfully!
   Total rows loaded: 19997

Loading C:\\Users\\decent\\OneDrive\\Desktop\\Excelerate_Project_Week1\\cleaned_CampaignData.csv into table 'campaign_data'...
campaign_data loaded successfully!
   Total rows loaded: 23

Loading C:\\Users\\decent\\OneDrive\\Desktop\\Excelerate_Project_Week1\\cleaned_OutreachData.csv into table 'outreach_data'...
outreach_data loaded successfully!
   Total rows loaded: 37435

 All datasets have been successfully loaded into PostgreSQL!
```



	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
8	351600	351600	351600	ratneshry06@gmail.com	ratneshry06@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
9	369415	369415	369415	dhruvvekariya.vmc18@gmail.com	dhruvvekariya.vmc18@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
10	345125	345125	345125	kanishk.9871@gmail.com	kanishk.9871@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
11	367981	367981	367981	gayathri.keenala@gmail.com	gayathri.keenala@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
12	349480	349480	349480	lakshaymunjal17@gmail.com	lakshaymunjal17@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
13	347104	347104	347104	shah.rahulsailesh@gmail.com	shah.rahulsailesh@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
14	360921	360921	360921	sekhar.fall2022@gmail.com	sekhar.fall2022@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
15	373583	373583	373583	glourdurani@gmail.com	glourdurani@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
16	351065	351065	351065	polishetty.jyothi08@gmail.com	polishetty.jyothi08@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
17	367276	367276	367276	chsrilatha.eduf23@gmail.com	chsrilatha.eduf23@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
18	362645	362645	362645	saimukeshjakkula2023@gmail.com	saimukeshjakkula2023@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
19	345987	345987	345987	phanendrakatta@gmail.com	phanendrakatta@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
20	346524	346524	346524	garv.career@gmail.com	garv.career@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
21	341882	341882	341882	likithapalakolanu08@gmail.com	likithapalakolanu08@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology
22	339396	339396	339396	saisandeep.pashem007@gmail.com	saisandeep.pashem007@gmail.com	Illinois Institute of Technology	[null]	Illinois Institute of Technology



	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
7	DNA24	DNA24	24	GR GS Call Campaign: India No Deposit	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
8	FA24AND	FA24AND	24	GR GS Call Campaign: Other ANF	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
9	FA24DNA	FA24DNA	24	GR GS Call Campaign: Other No Deposit	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
10	FA24DNI	FA24DNI	24	GR GS SP25 Campaign- All I-20s Sent	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
11	FA24IP	FA24IP	24	GR GS SP25 Campaign- Admit No Deposit	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
12	FA24SIC	FA24SIC	24	GR GS SP25 Campaign- Deposit No I-20	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
13	IANF23	IANF23	23	GR GS SP25 Campaign- Deferrals to SP25	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
14	IND23	IND23	23	GR GS SP25 Campaign- In Progress	[null]	Pre Admission	Pre Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
15	OANF23	OANF23	23	GR GS SP25 Campaign- New Inquiry	[null]	Pre Admission	Pre Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
16	OND23	OND23	23	GR GS SP25 Campaign- Submitted Incomplete	[null]	Pre Admission	Pre Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
17	SP25AI2S	SP25AI2S	28	GR GS Call Campaign: Africa ANF	[null]	Pre Admission	Pre Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
18	SP25AND	SP25AND	25	GR GS Call Campaign: Africa No Deposit	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
19	SP25DN1	SP25DN1	28	GR GS Call Campaign: Bangladesh Pakistan Nepal ANF	[null]	Pre Admission	Pre Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
20	SP25DSP	SP25DSP	25	GR GS Call Campaign: Bangladesh Pakistan Nepal No Deposit	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
21	SP25IP	SP25IP	25	GR GS Call Campaign: China Taiwan Korea ANF	[null]	Pre Admission	Pre Admission	AY2024	AY2024	2024	Illinois Institute of Technolog
22	SP25NIQ	SP25NIQ	25	Deposit and Advised Not Enrolled	[null]	Post Admission	Post Admission	AY2024	AY2024	2024	Illinois Institute of Technolog



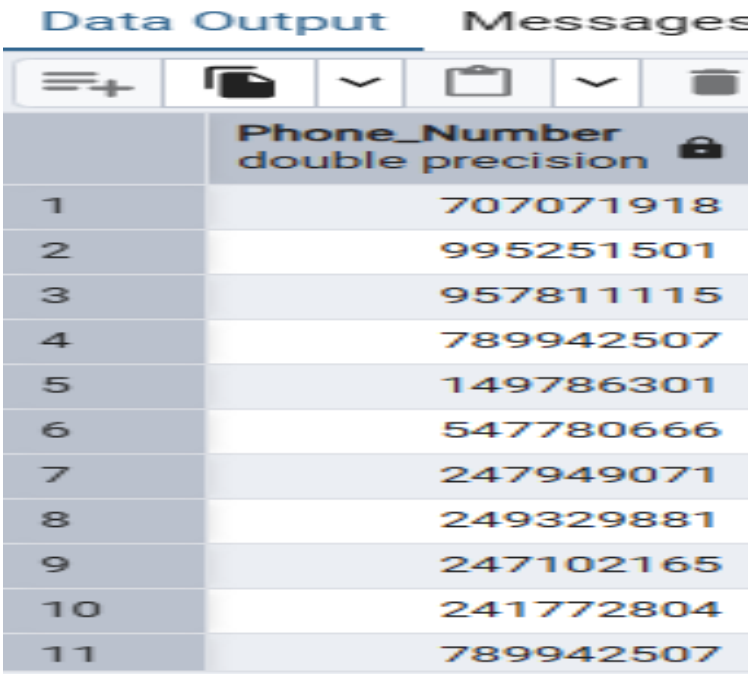
6. Running and Performing Queries

The following queries were executed in PostgreSQL to perform data analysis on the cleaned datasets.

- **Top 10 Countries:** This query counts the number of applicants from each country and sorts them to show the top 10 with the most applicants, providing insight into the geographical distribution of the applicant pool.

	Country text	applicant_count bigint
1	N/A	19380
2	mohanendraapplication@gmail.com	2
3	neda.saffari7115@gmail.com	1
4	nitya2411@gmail.com	1
5	rifatrahman097@gmail.com	1
6	prathyush.lebaku7@gmail.com	1
7	gunsai22@gmail.com	1
8	nshrey53@gmail.com	1
9	surlekarh@gmail.com	1
10	pradeepbandla123@gmail.com	1

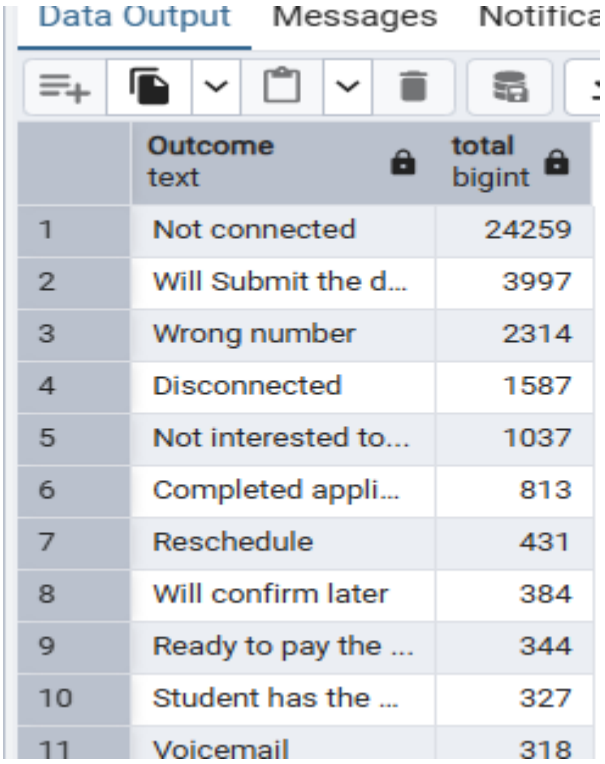
- **Applicants with Invalid Phone Numbers:** This query identifies applicants whose phone numbers do not have the expected 10-digit length, which helps to pinpoint data inconsistencies.



Data Output Messages

	Phone_Number double precision
1	707071918
2	995251501
3	957811115
4	789942507
5	149786301
6	547780666
7	247949071
8	249329881
9	247102165
10	241772804
11	789942507

- **Joining Campaign and Outreach:** By joining the outreach_data and campaign_data tables, this query counts the total records for each outreach outcome (e.g., converted, not converted), providing a high-level view of overall results.



Data Output Messages Notifica

	Outcome text	total bigint
1	Not connected	24259
2	Will Submit the d...	3997
3	Wrong number	2314
4	Disconnected	1587
5	Not interested to...	1037
6	Completed appli...	813
7	Reschedule	431
8	Will confirm later	384
9	Ready to pay the ...	344
10	Student has the ...	327
11	Voicemail	318

- **Count of Outcomes Per Campaign:** This query gives a more detailed breakdown by counting outcomes for each specific campaign, allowing you to assess the performance of individual campaigns.

Data Output | Messages | Notifications

	Name text	Outcome text	total bigint
1	Deposit and Advised Not Enrolled	Not connected	2100
2	Deposit and Advised Not Enrolled	Wrong number	1220
3	Deposit and Advised Not Enrolled	Disconnected	201
4	Deposit and Advised Not Enrolled	Will start applic...	199
5	Deposit and Advised Not Enrolled	Still making a de...	41
6	Deposit and Advised Not Enrolled	Application alrea...	36
7	Deposit and Advised Not Enrolled	Will Submit the d...	27
8	Deposit and Advised Not Enrolled	Completed appli...	19
9	Deposit and Advised Not Enrolled	Not interested to...	16
10	Deposit and Advised Not Enrolled	Voicemail	12
11	Deposit and Advised Not Enrolled	Reschedule	11

Total rows: 289 Query complete 00:00:00.230

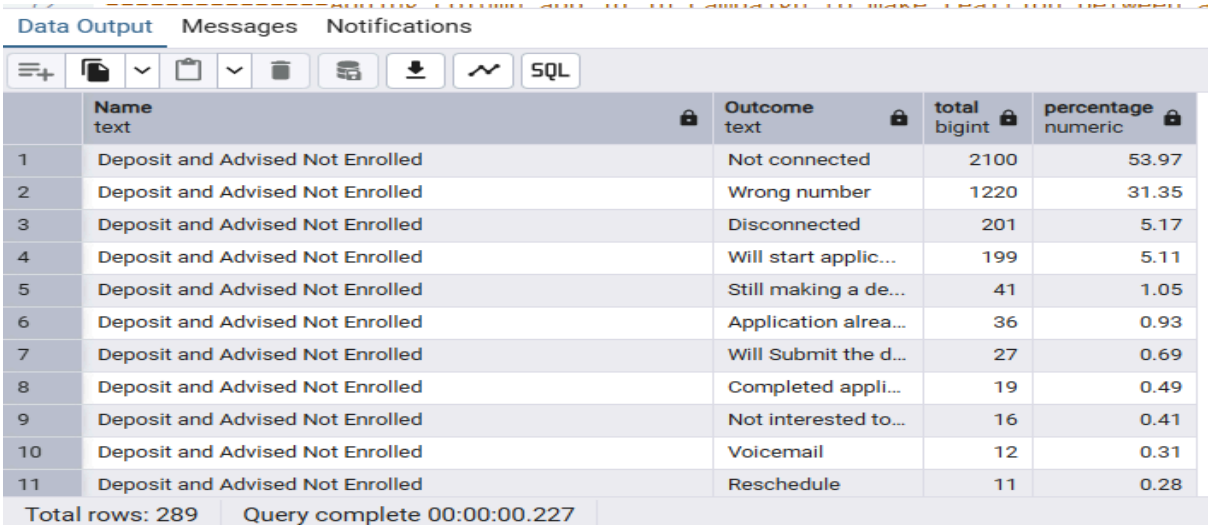
- **Total Outreach Per Campaign:** This query calculates the total outreach volume for each campaign, showing which campaigns had the most activity.

Data Output | Messages | Notifications

	Name text	total_outreach bigint
1	GR GS SP25 Campaign- Admit No Deposit	9603
2	GR GS SP25 Campaign- Deposit No I-20	5640
3	Deposit and Advised Not Enrolled	3891
4	GR GS Call Campaign: China Taiwan Korea ANF	3667
5	GR GS SP25 Campaign- Deferrals to SP25	2697
6	GR GS Call Campaign: Other ANF	1846
7	Deposit Not Advised	1838
8	GR GS Call Campaign: Bangladesh Pakistan Nepal No Deposit	1229
9	GR GS SP25 Campaign- In Progress	1168
10	GR GS Call Campaign: Africa ANF	948
11	GR GS SP25 Campaign- All I-20s Sent	894

Total rows: 23 Query complete 00:00:00.128

- **Percentage of Each Outcome Per Campaign:** This advanced query determines the percentage of each outcome within a campaign, enabling a standardized comparison of success rates across different campaigns.

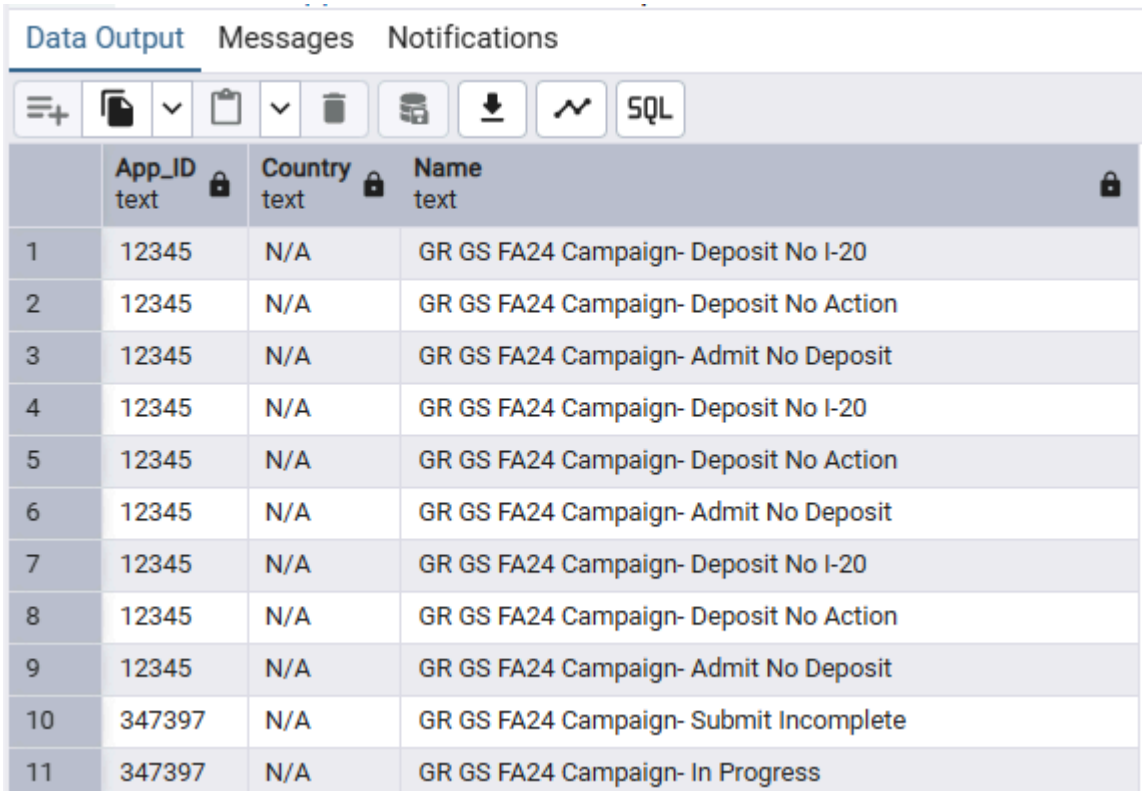


Data Output | Messages | Notifications

	Name text	Outcome text	total bigint	percentage numeric
1	Deposit and Advised Not Enrolled	Not connected	2100	53.97
2	Deposit and Advised Not Enrolled	Wrong number	1220	31.35
3	Deposit and Advised Not Enrolled	Disconnected	201	5.17
4	Deposit and Advised Not Enrolled	Will start applic...	199	5.11
5	Deposit and Advised Not Enrolled	Still making a de...	41	1.05
6	Deposit and Advised Not Enrolled	Application alrea...	36	0.93
7	Deposit and Advised Not Enrolled	Will Submit the d...	27	0.69
8	Deposit and Advised Not Enrolled	Completed appli...	19	0.49
9	Deposit and Advised Not Enrolled	Not interested to...	16	0.41
10	Deposit and Advised Not Enrolled	Voicemail	12	0.31
11	Deposit and Advised Not Enrolled	Reschedule	11	0.28

Total rows: 289 Query complete 00:00:00.227

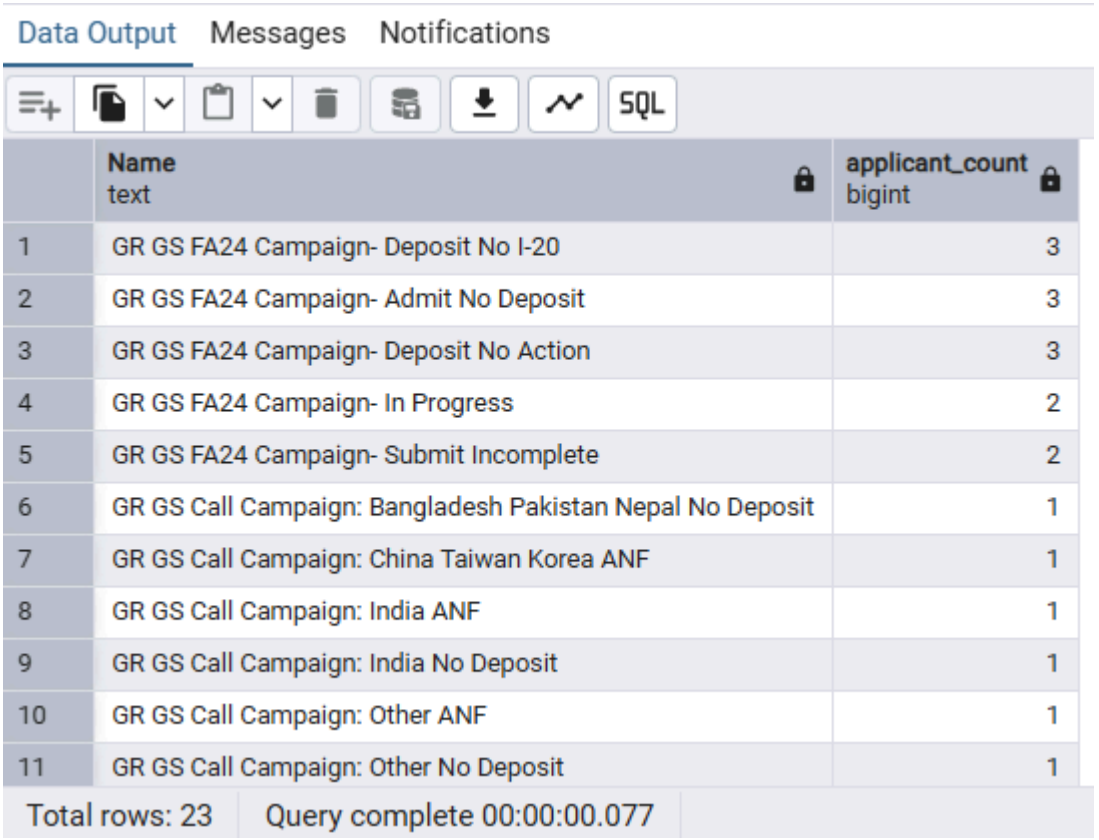
- **Applicants with Their Campaigns:** This query joins the applicant_data and campaign_data tables to list each applicant and their assigned campaign, linking individuals to specific initiatives.



Data Output | Messages | Notifications

	App_ID text	Country text	Name text
1	12345	N/A	GR GS FA24 Campaign- Deposit No I-20
2	12345	N/A	GR GS FA24 Campaign- Deposit No Action
3	12345	N/A	GR GS FA24 Campaign- Admit No Deposit
4	12345	N/A	GR GS FA24 Campaign- Deposit No I-20
5	12345	N/A	GR GS FA24 Campaign- Deposit No Action
6	12345	N/A	GR GS FA24 Campaign- Admit No Deposit
7	12345	N/A	GR GS FA24 Campaign- Deposit No I-20
8	12345	N/A	GR GS FA24 Campaign- Deposit No Action
9	12345	N/A	GR GS FA24 Campaign- Admit No Deposit
10	347397	N/A	GR GS FA24 Campaign- Submit Incomplete
11	347397	N/A	GR GS FA24 Campaign- In Progress

- **Count of Applicants Per Campaign:** This final query counts the number of applicants associated with each campaign, providing a simple metric for campaign engagement.



Data Output | Messages | Notifications

	Name text	applicant_count bigint
1	GR GS FA24 Campaign- Deposit No I-20	3
2	GR GS FA24 Campaign- Admit No Deposit	3
3	GR GS FA24 Campaign- Deposit No Action	3
4	GR GS FA24 Campaign- In Progress	2
5	GR GS FA24 Campaign- Submit Incomplete	2
6	GR GS Call Campaign: Bangladesh Pakistan Nepal No Deposit	1
7	GR GS Call Campaign: China Taiwan Korea ANF	1
8	GR GS Call Campaign: India ANF	1
9	GR GS Call Campaign: India No Deposit	1
10	GR GS Call Campaign: Other ANF	1
11	GR GS Call Campaign: Other No Deposit	1

Total rows: 23 | Query complete 00:00:00.077

---

7. Conclusion

The EDA process successfully prepared three raw datasets for analysis. The cleaning pipeline effectively handled missing values, standardized data formats, and removed duplicates, resulting in a cleaner and more reliable set of data. The cleaned data was then successfully loaded into a PostgreSQL database, setting the stage for further database-centric analysis.