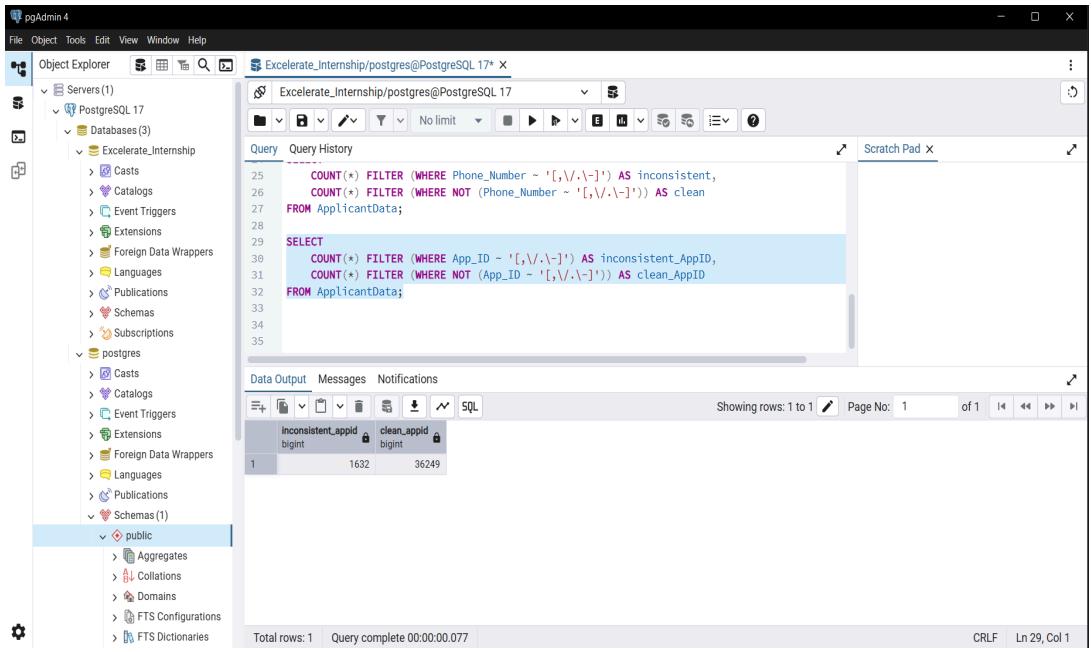
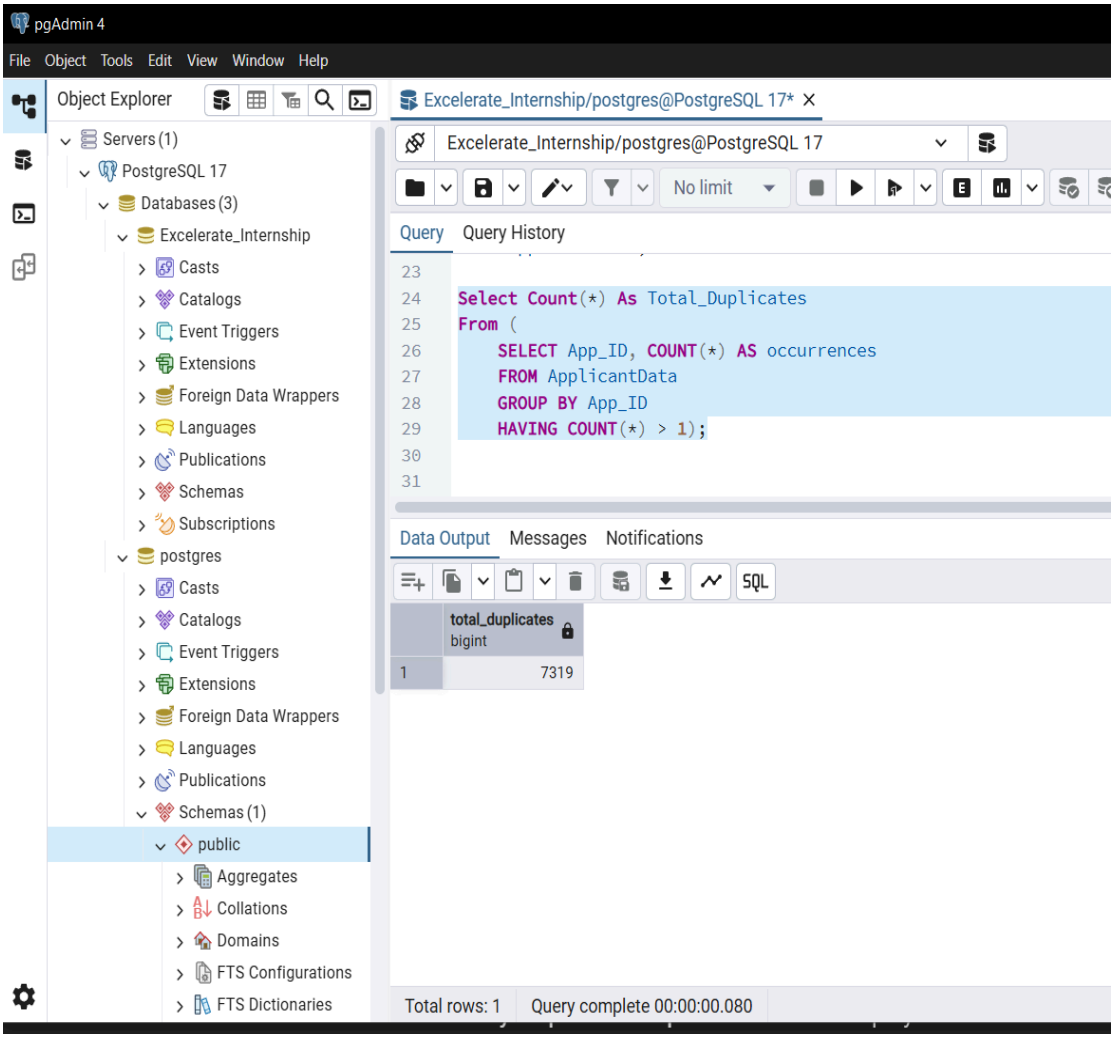
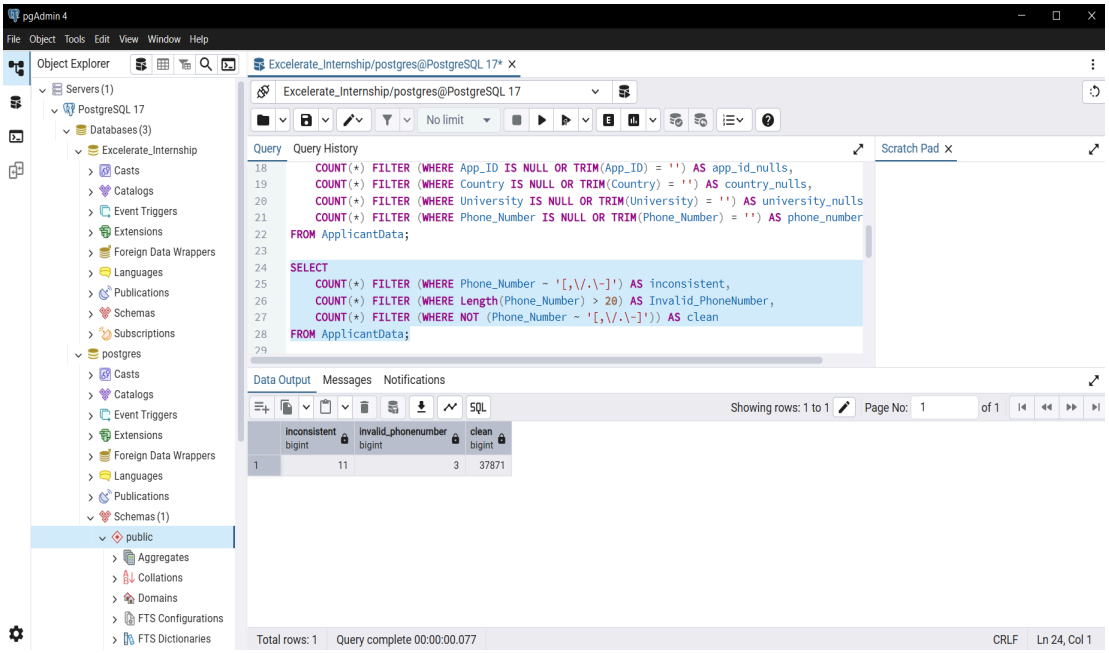
Week_1 Report

1. Exploratory Data Analysis:

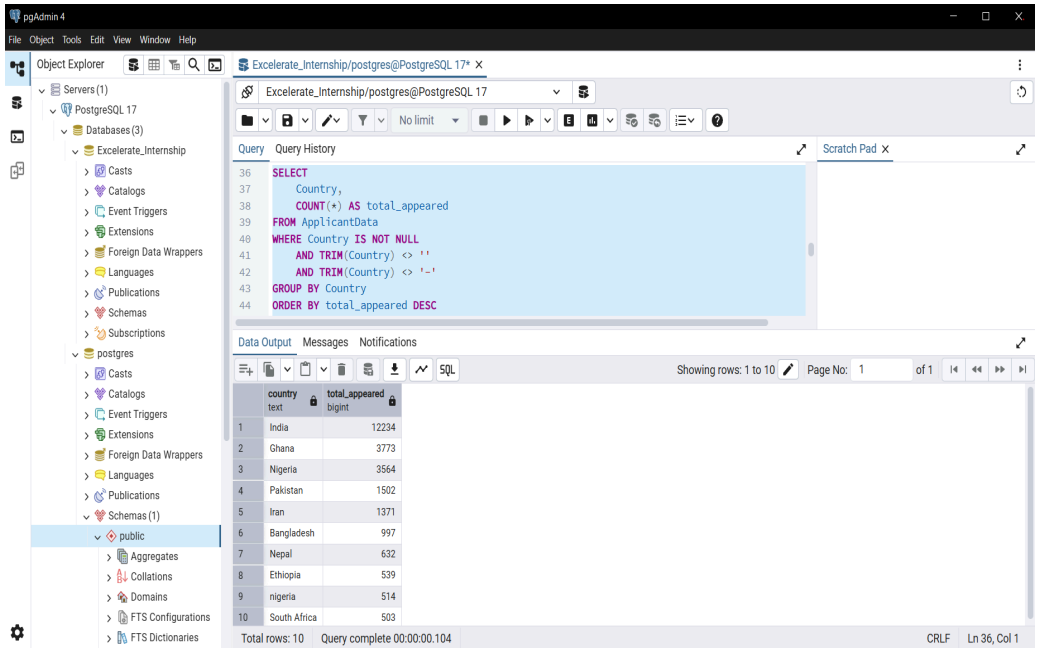
- **Applicant Data:**
  - Applicant Data dataset had 37883 rows with 4 columns.
  - The column names are App_ID, Country, University, Phone_Number. Every column of this file was in general format.
  - 

- 

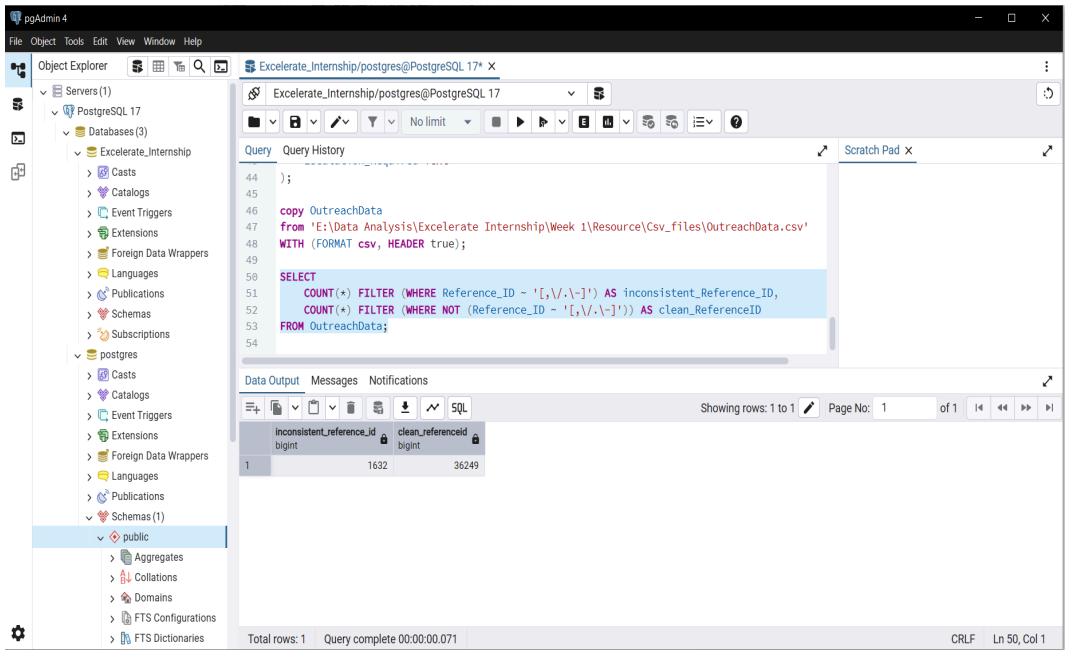
- From here we can see, App_ID has 1632 inconsistent data with 7319 duplicates which is polluting the data integrity. Also 36249 clean data.



- Here we can see, 11 inconsistent phone_number with 3 invalid phone numbers. Moreover, it has 37871 clean data.
- University column has only 1 name which is “Illinois Institute of Technology”. So it is evident that the data is of this university or related to this university.
- Top 10 most countries people applied from are India, Ghana, Nigeria, Pakistan, Iran, Bangladesh, Nepal, Ethiopia, Nigeria, South Africa.



- **Campaign Data:**
  - Applicant Data dataset had 24 rows with 7 columns.
  - The column names are ID, Name, Category, Intake, University, Status, and Start_Date. Every column of this file was in general format.
- **Outreach Data:**
  - Applicant Data dataset had 37882 rows with 8 columns.
  - The column names are ID, Country, University, Phone_Number. Every column of this file was in general format.
  - OutreachData's Reference_ID column has 1632 inconsistent data.



2. Data Cleaning:

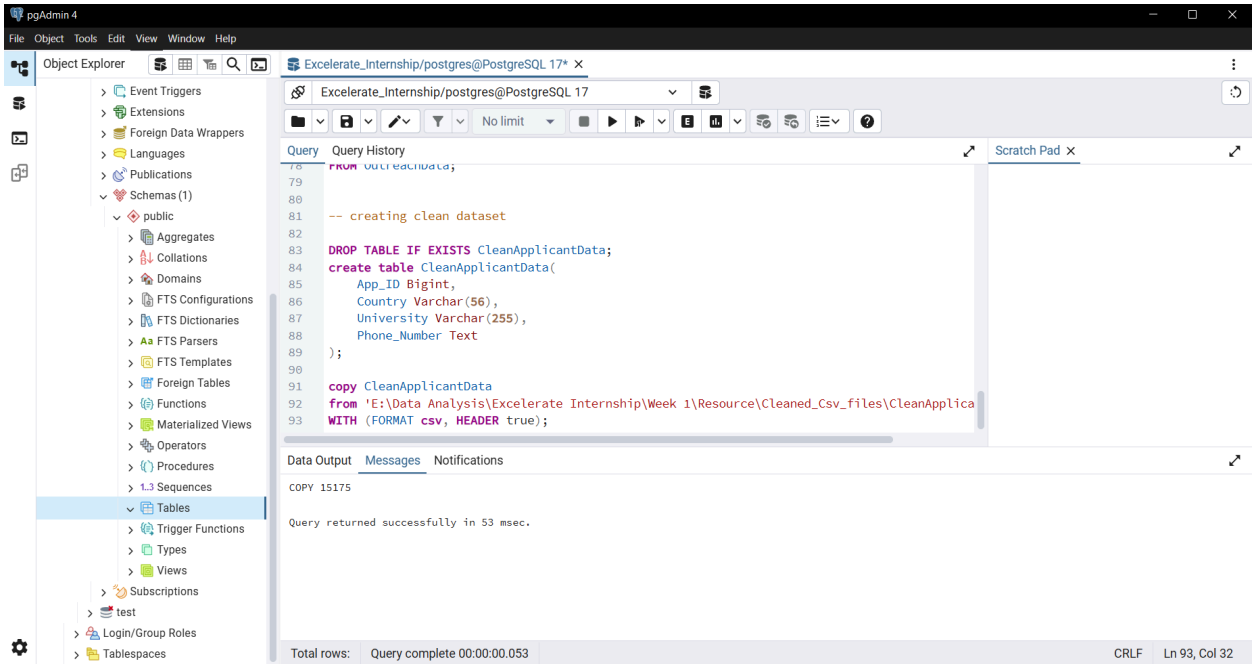
- **Applicant Data:**
    i. In this dataset, under “App_ID” column which is the primary key of this dataset, has null values, inconsistent values and missing values. To solve these issues, we have dropped all the rows where primary key is missing. And then found inputs like combinations of (`, ….. , //, “) these symbols. So, we deleted the corresponding rows which have these symbols. Moreover, we faced that there were some text data and primary key can not be text data which makes us to drop those rows too.
    ii. Country Column has lot of inappropriate values like emails, normal text and which will create an bad effect. And that is the reason I filtered out all the input not having country names and changed those inputs to “Not Mentioned” so we kept the data integrity by not dropping rows.
    iii. Phone_Number column was on “General” format which could be problematic for data modeling and is the reason why we changed it to number format.
- **Campaign Data:**
    i. In this dataset, there was a Start_Date column which we changed the format to Date.

ii. Additionally, there was a column called “name” which had informations like season, campaign type, region, and status. These informations are very necessary for the visualizations part if we can categorize them. So, we created 4 new column and named them as Season, Campaign_Region, Campaign_Type, and Campaign_Status. After that stored necessary informations into these newly created columns and if there was a nothing to put in respective column used “Not Mentioned” to avoid future sufferings when we will work on visualization.

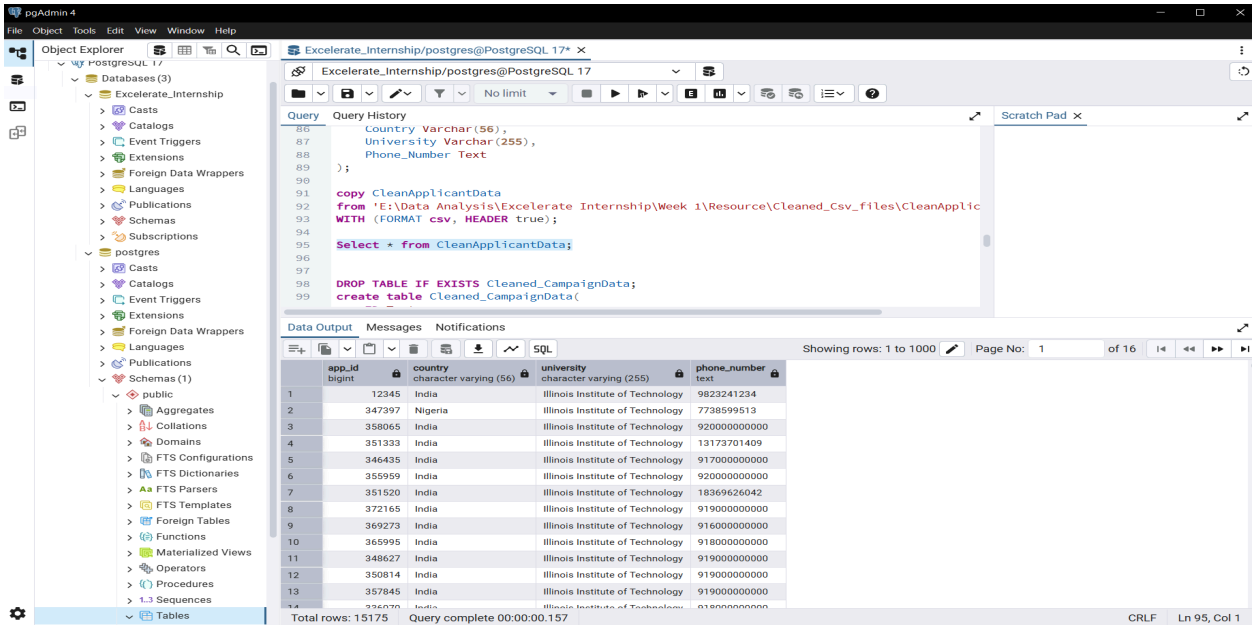
- **Outreach Data:**
  i. “OutreachData” dataset had Reference_ID which was connected to “App_ID” of applicant data. But the difference is outreach data can have multiple same “Reference_ID” as same applicant can submit their data multiple times. So, here we kept all the duplicate values. But similar to “App_ID” cleaning process, we dropped all the rows having missing values and inconsistent values.
  ii. The dataset also had a crucial column called “Received_At” from which we created 6 new columns called Date, Year, Time, Hour, Weekday, and Month and added informations to these columns from “Received_At” column for the better visualizations and finding trends, patterns from the data.
  iii. Also added a “Outreach_ID” column which will be the primary key in this dataset and this will help in the data modeling like connecting with other dataset/tables in PostgreSQL for finding trends and patterns out of the data.

3. PostgreSql Setup Proof:

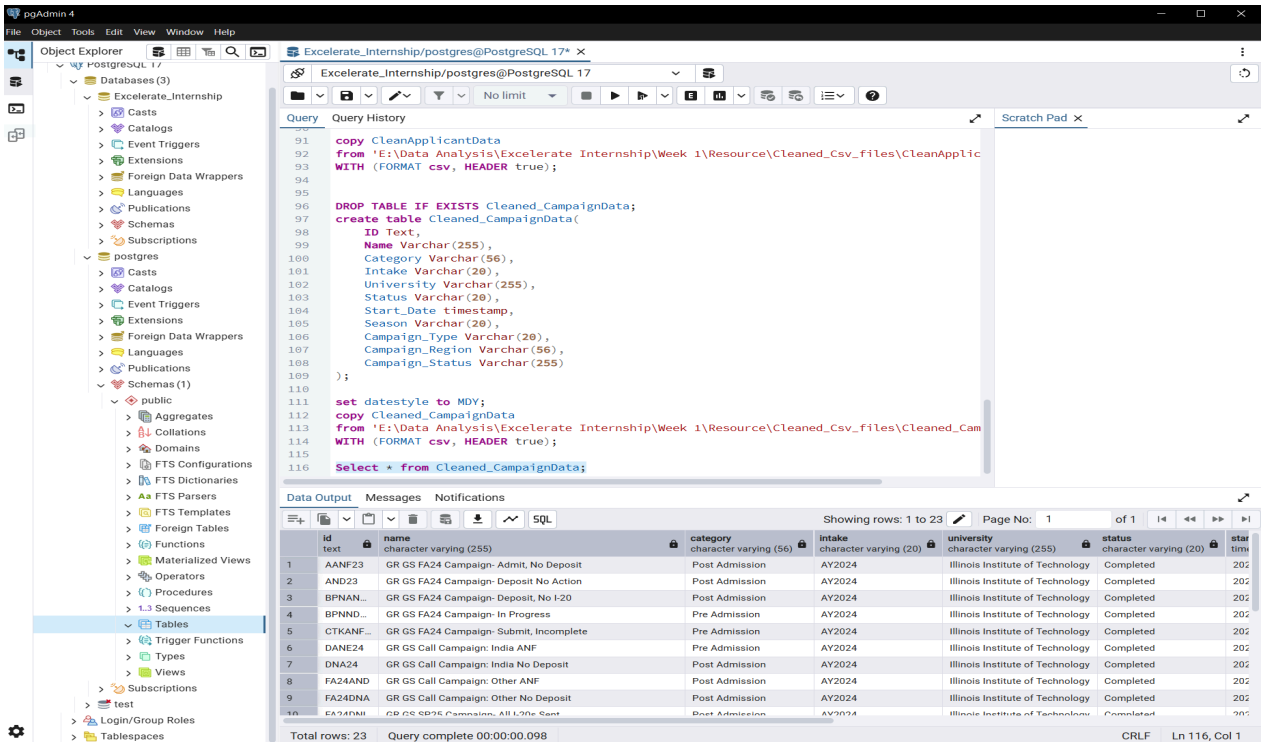
- Installed PostgreSql in my system. Importing Cleaned dataset into the PostgreSql to find trends and patterns from the data.



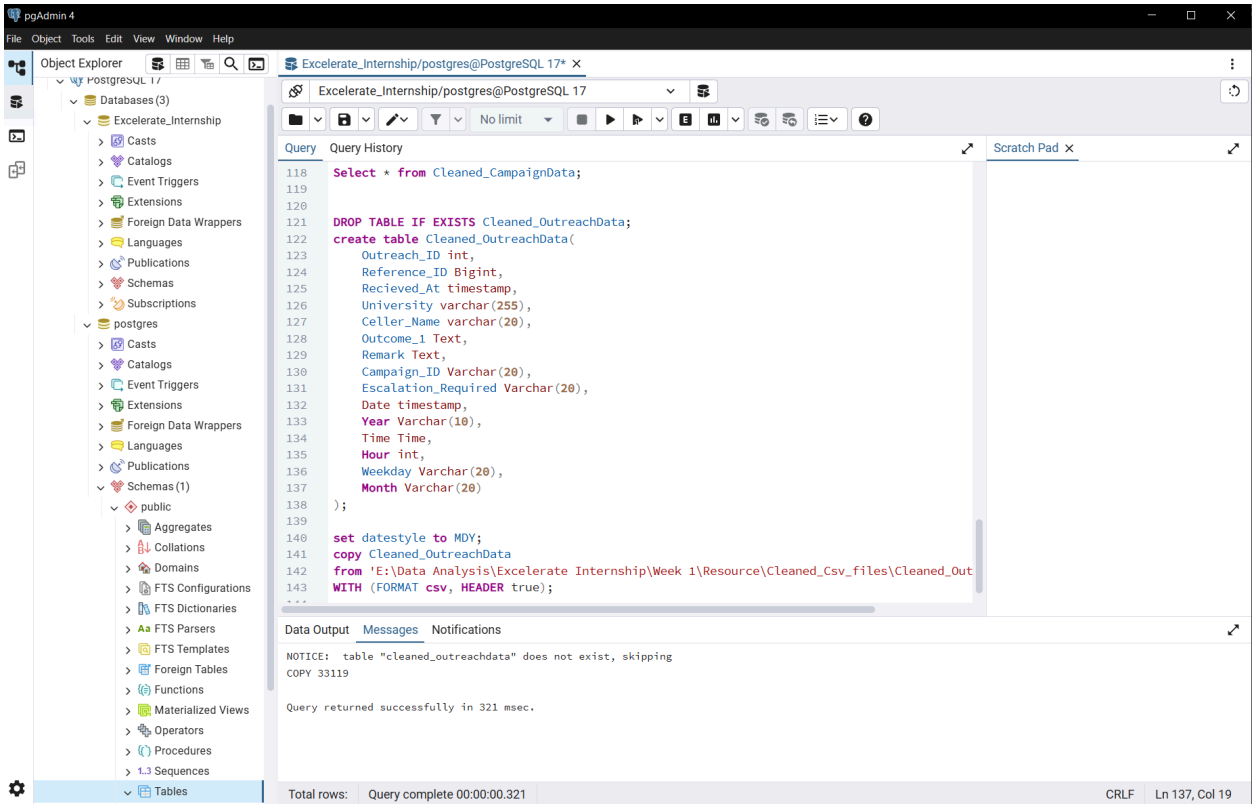
View of connected Cleaned Applicant Data.

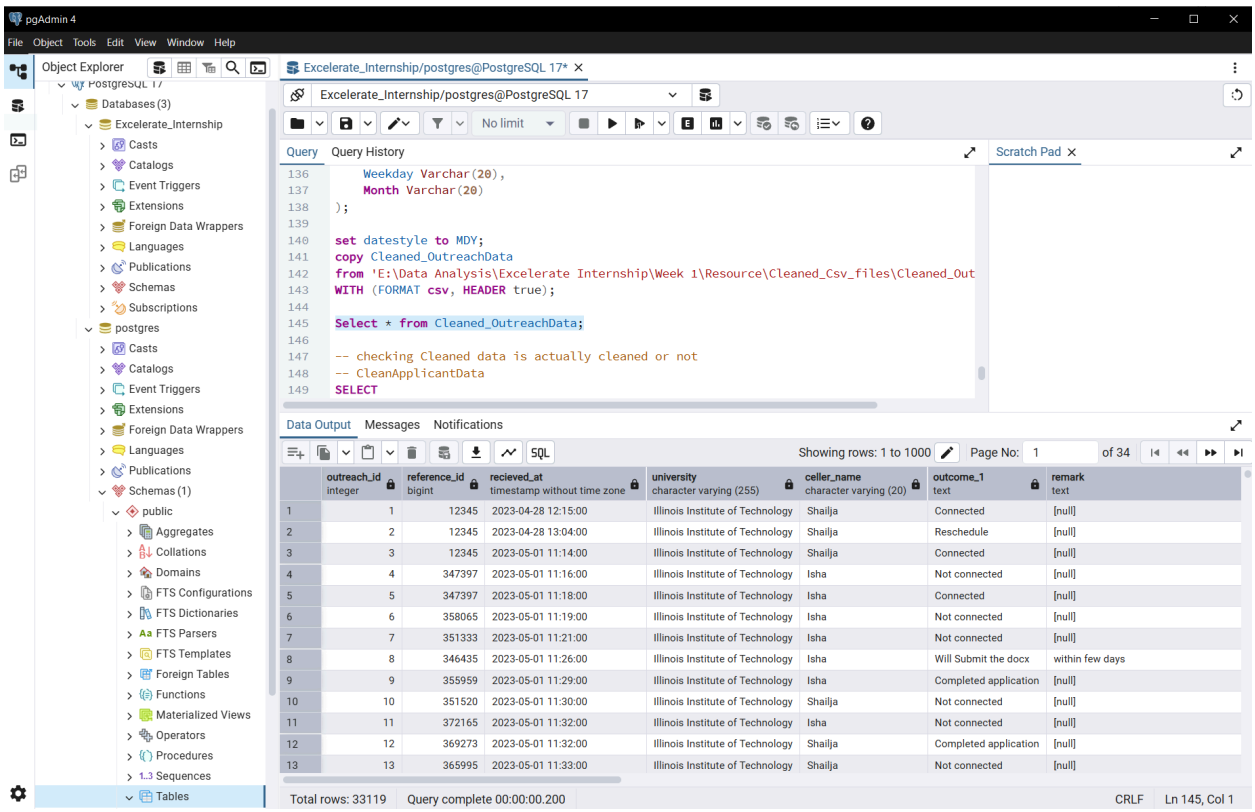


Also connected the Cleaned_CampaignData to postgreSql.



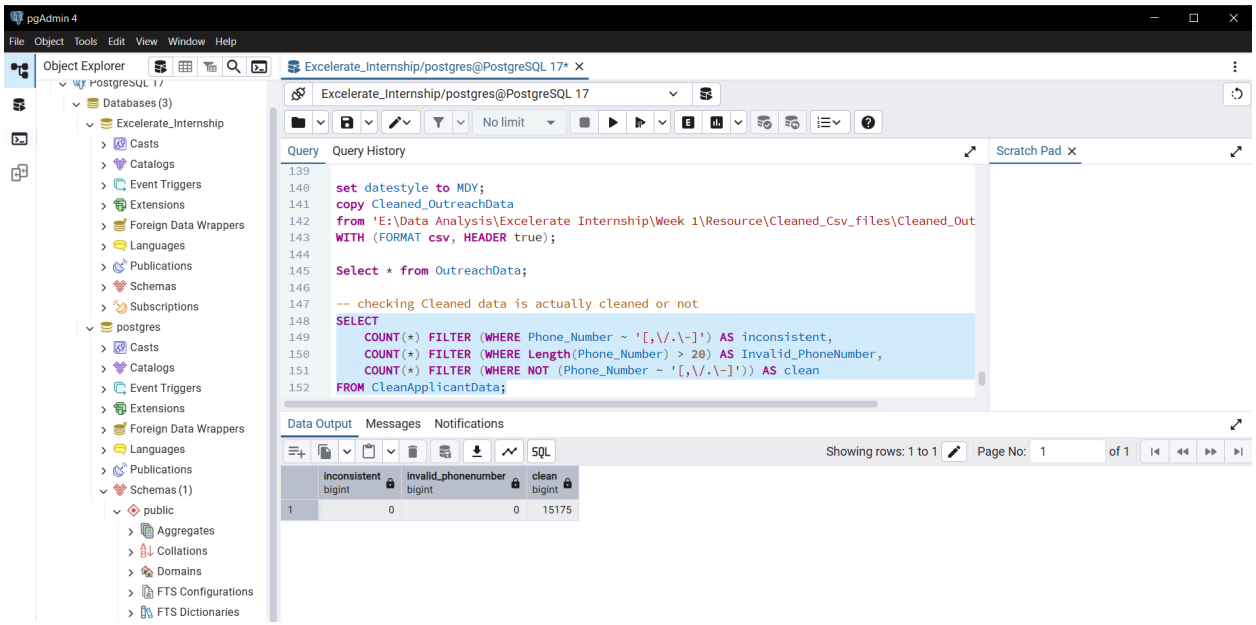
Cleaned_OutreachData is alson ready to find trends and patterns using PostgreSql.



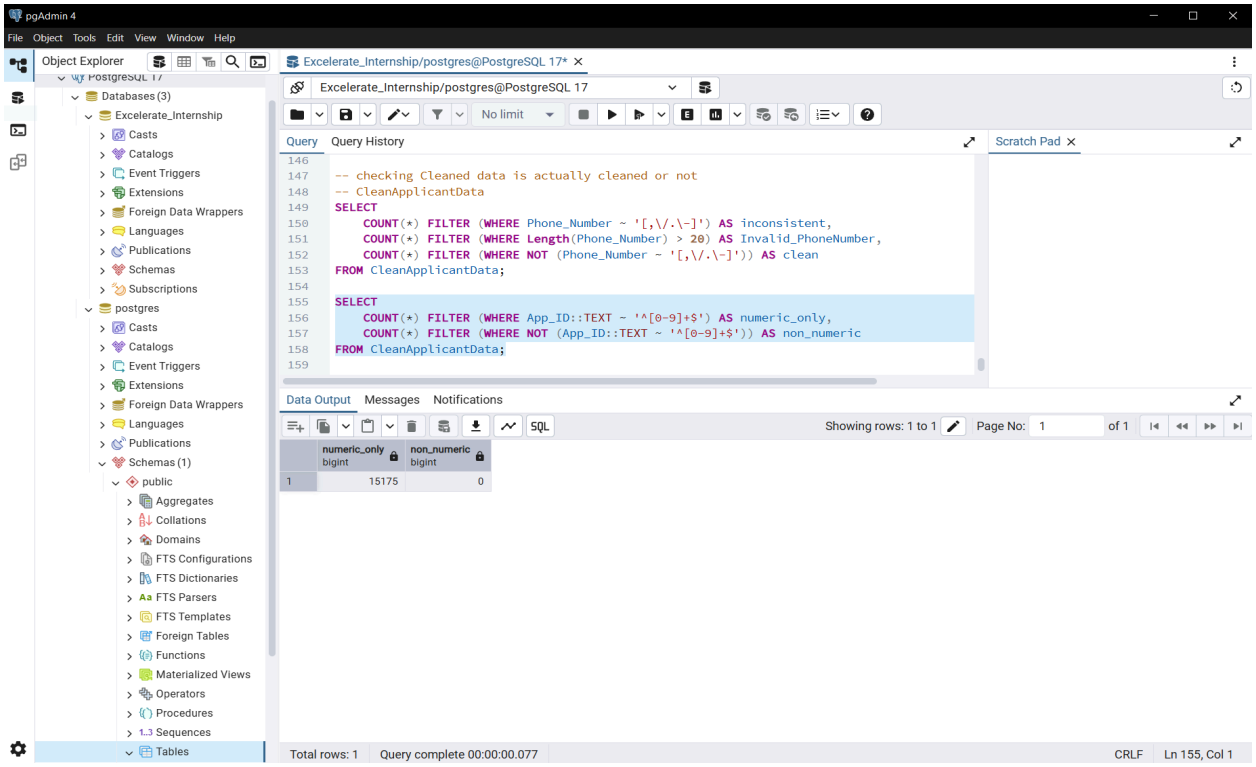


	outreach_id integer	reference_id bigint	recieved_at timestamp without time zone	university character varying (255)	celler_name character varying (20)	outcome_1 text	remark text
1	1	12345	2023-04-28 12:15:00	Illinois Institute of Technology	Shailja	Connected	[null]
2	2	12345	2023-04-28 13:04:00	Illinois Institute of Technology	Shailja	Reschedule	[null]
3	3	12345	2023-05-01 11:14:00	Illinois Institute of Technology	Shailja	Connected	[null]
4	4	347397	2023-05-01 11:16:00	Illinois Institute of Technology	Isha	Not connected	[null]
5	5	347397	2023-05-01 11:18:00	Illinois Institute of Technology	Isha	Connected	[null]
6	6	358065	2023-05-01 11:19:00	Illinois Institute of Technology	Isha	Not connected	[null]
7	7	351333	2023-05-01 11:21:00	Illinois Institute of Technology	Isha	Not connected	[null]
8	8	346435	2023-05-01 11:26:00	Illinois Institute of Technology	Isha	Will Submit the docx	within few days
9	9	355959	2023-05-01 11:29:00	Illinois Institute of Technology	Isha	Completed application	[null]
10	10	351520	2023-05-01 11:30:00	Illinois Institute of Technology	Shailja	Not connected	[null]
11	11	372165	2023-05-01 11:32:00	Illinois Institute of Technology	Isha	Not connected	[null]
12	12	369273	2023-05-01 11:32:00	Illinois Institute of Technology	Shailja	Completed application	[null]
13	13	365995	2023-05-01 11:33:00	Illinois Institute of Technology	Shailja	Not connected	[null]

- **Exploring the “Clean_ApplicantData”:**

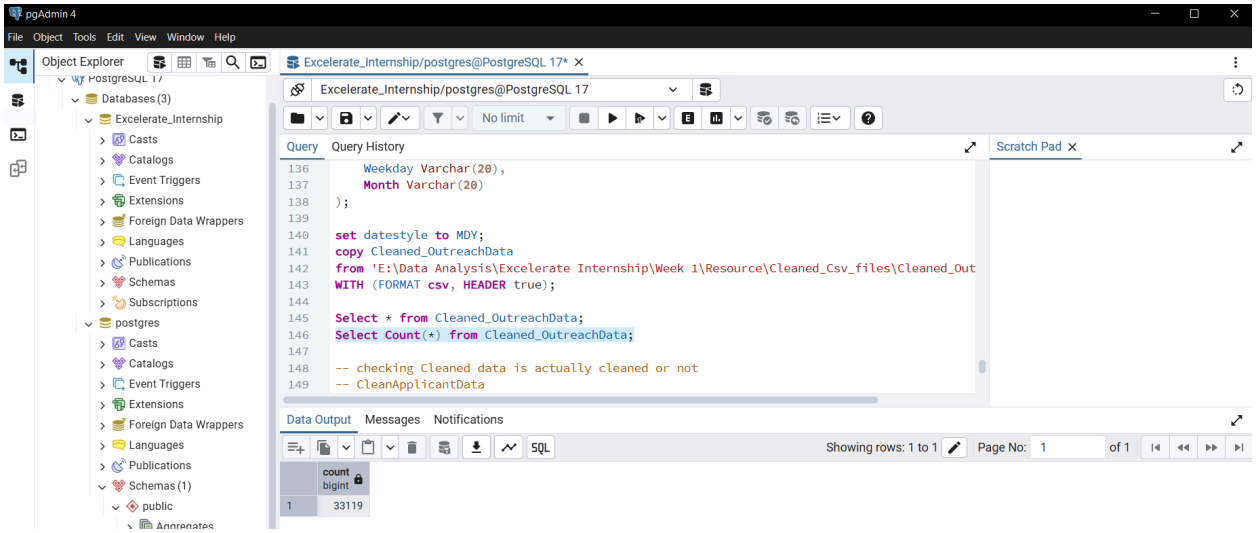


No inconsistent data is found in “Phone_Number” column of cleaned applicant data.

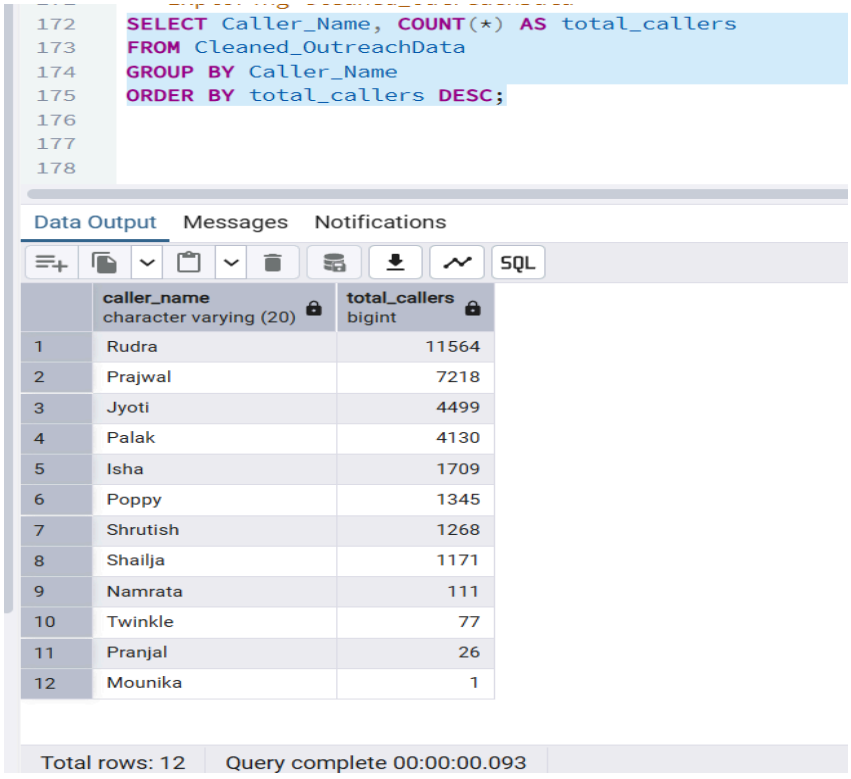


So, it is clear that there is not any non numeric value in our primary key which is the most important column, “App_ID”.

- **Exploring the “Cleaned_OutreachData”:**



These are the callers who has been contacting with applicants.



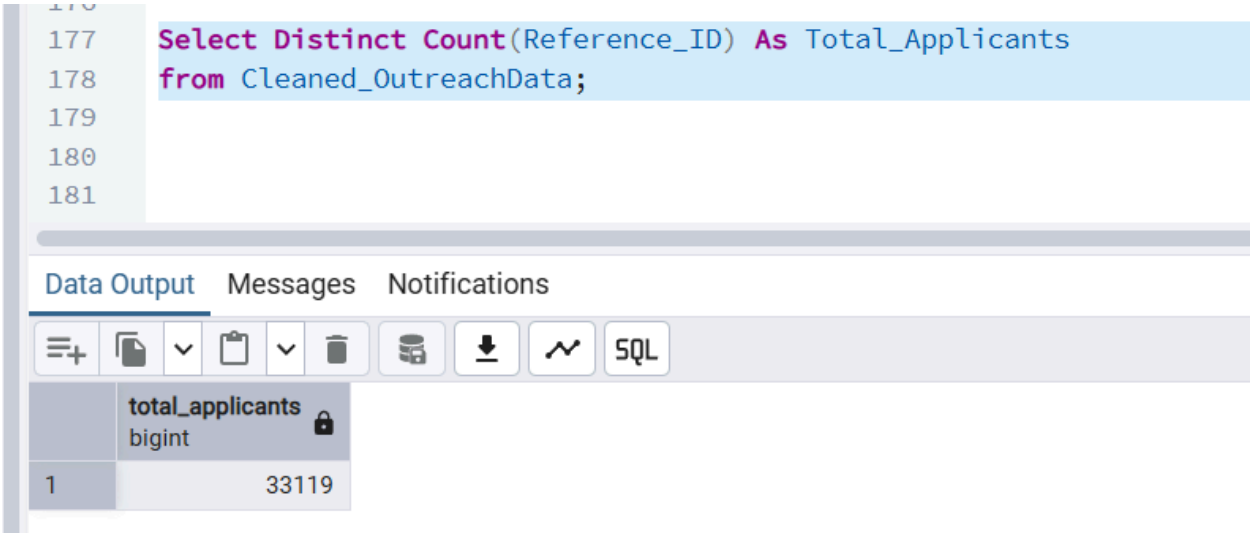
```
SELECT Caller_Name, COUNT(*) AS total_callers
FROM Cleaned_OutreachData
GROUP BY Caller_Name
ORDER BY total_callers DESC;
```

Data Output | Messages | Notifications

	caller_name character varying (20)	total_callers bigint
1	Rudra	11564
2	Prajwal	7218
3	Jyoti	4499
4	Palak	4130
5	Isha	1709
6	Poppy	1345
7	Shrutish	1268
8	Shailja	1171
9	Namrata	111
10	Twinkle	77
11	Pranjal	26
12	Mounika	1

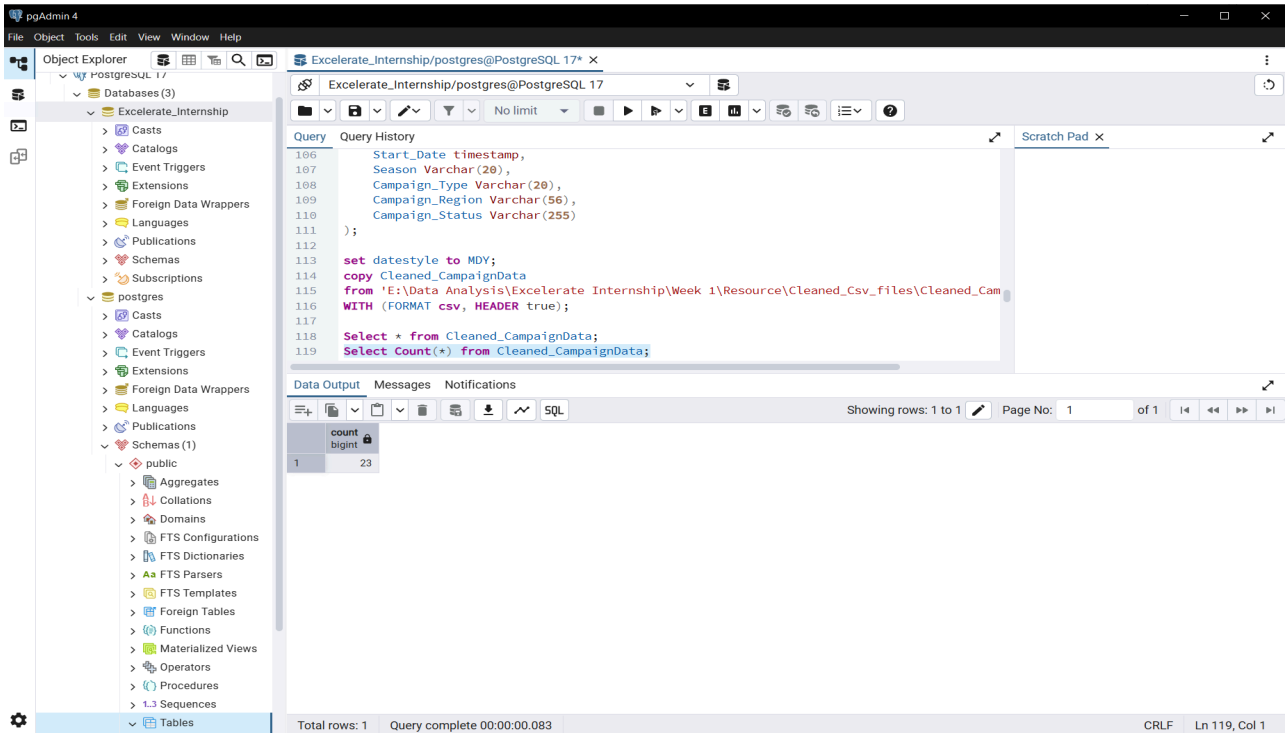
Total rows: 12 Query complete 00:00:00.093

From the above picture we can see, Rudra has the most clients with 11564. Prjawal and Jyoti has the second and third most clients. So, we can say, Rudra is the top performer among the callers.

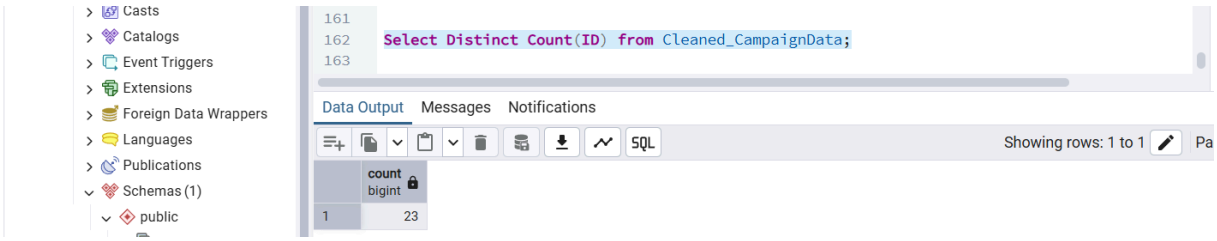


From Cleaned_OutreachData, we get to know there are 33119 applicants applied in “Illinois Institute of Technology” University through different campaigns.

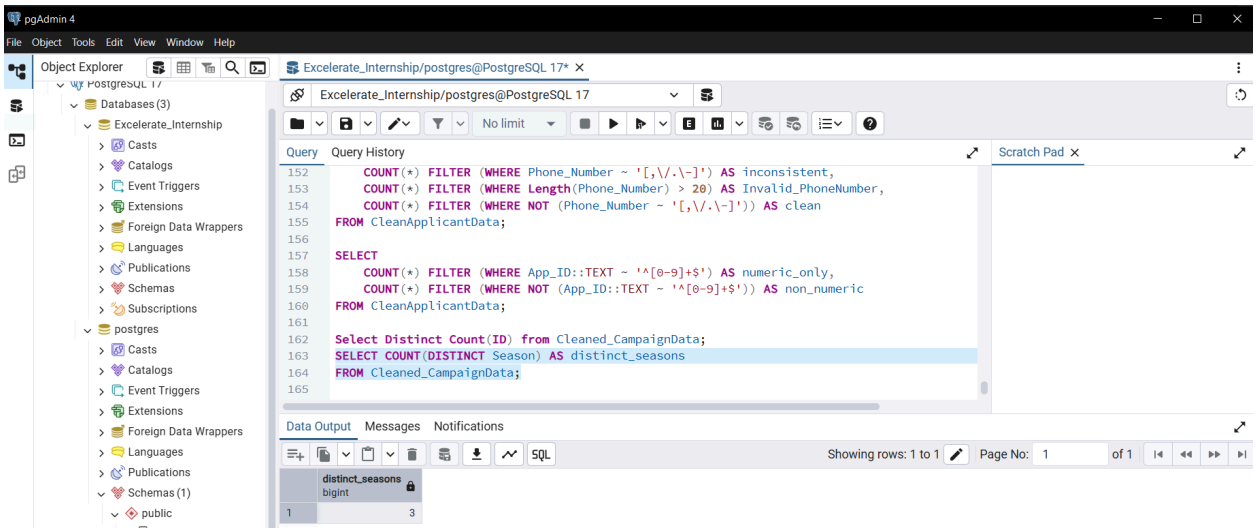
- **Exploring the “Cleaned_CampaignData”**



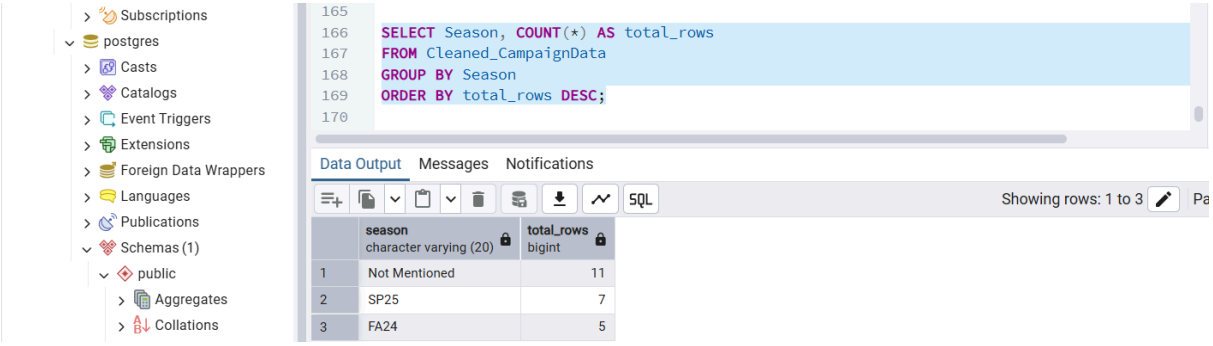
All the rows are there in the dataset after the the integration with Postgresql.



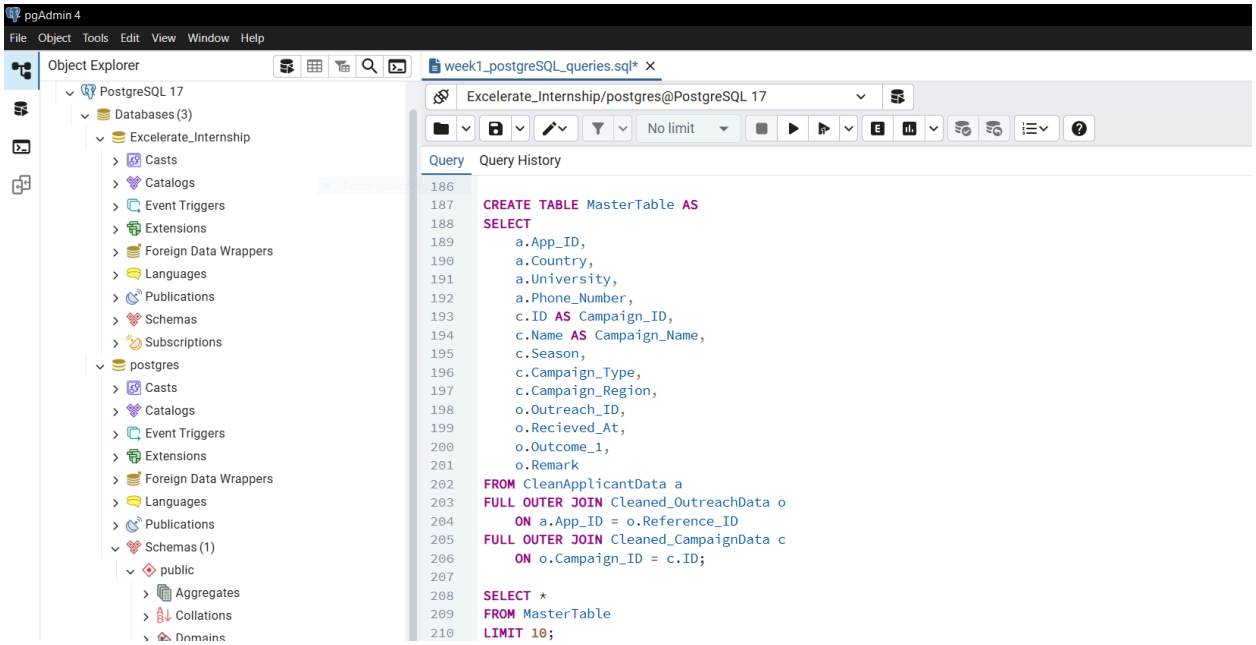
The primary key of Cleaned_CampaignData is all unique as we got same row nums.

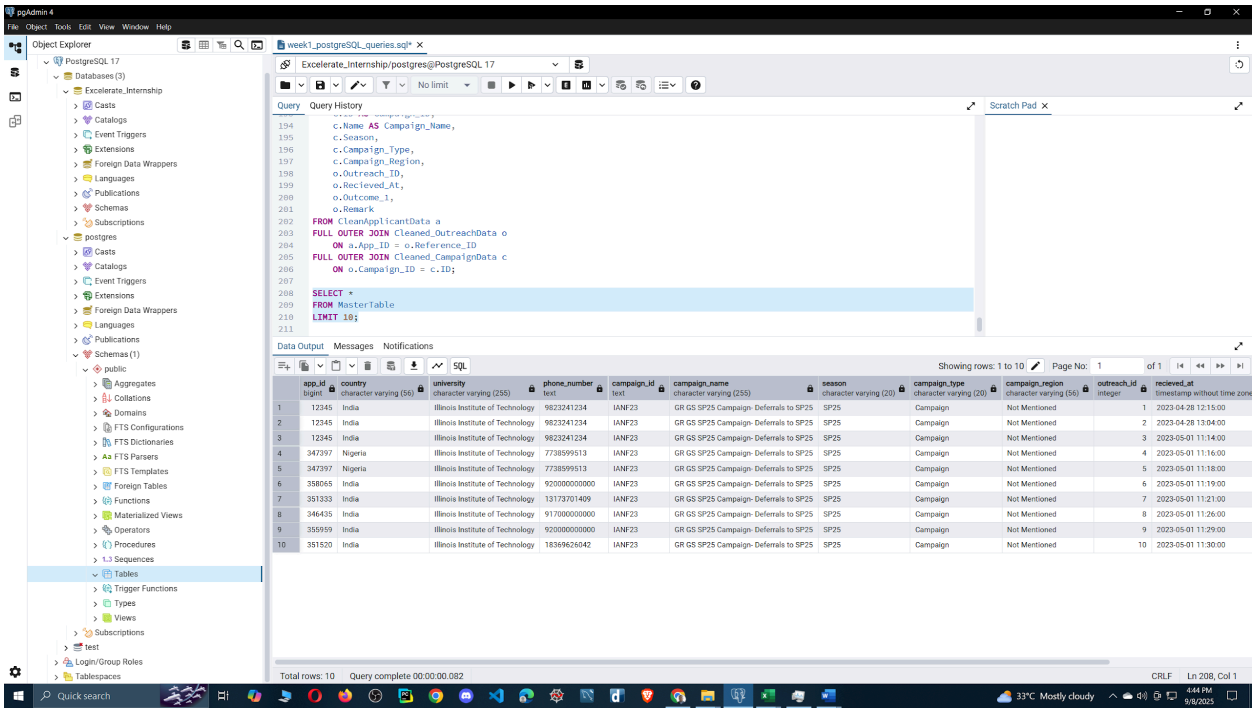


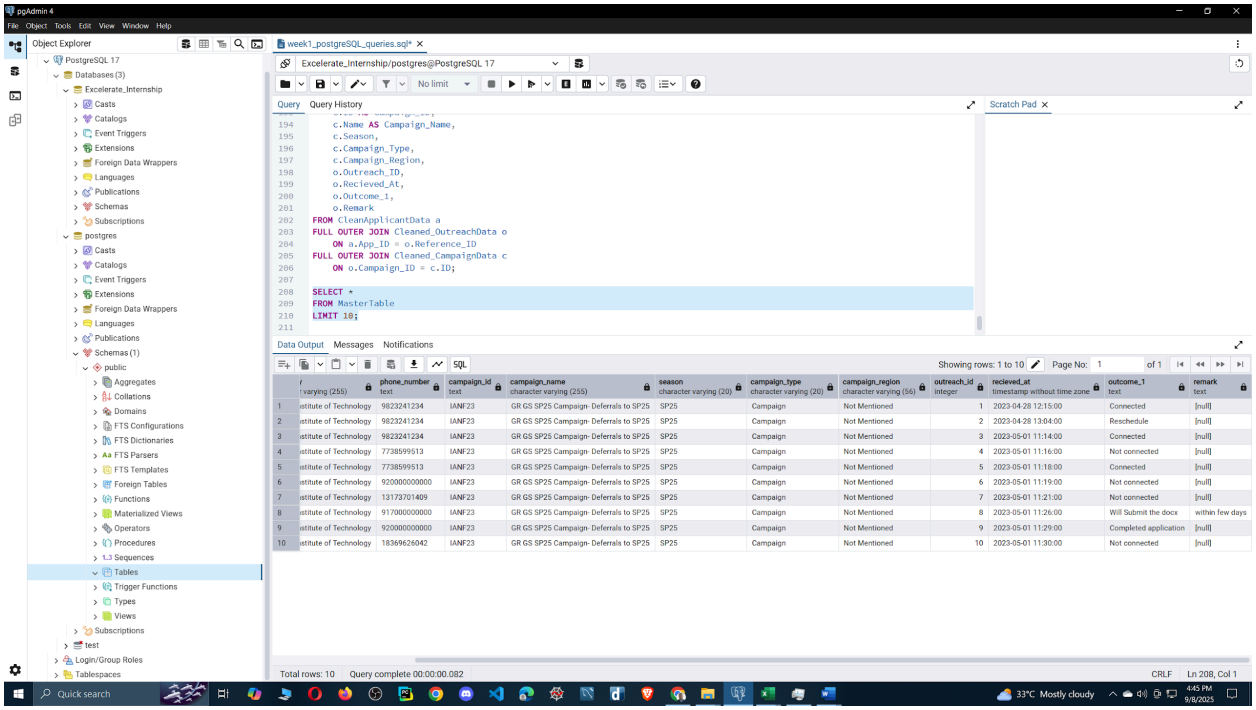
There are 3 types of entries in the season column. And These are the entries of this column.

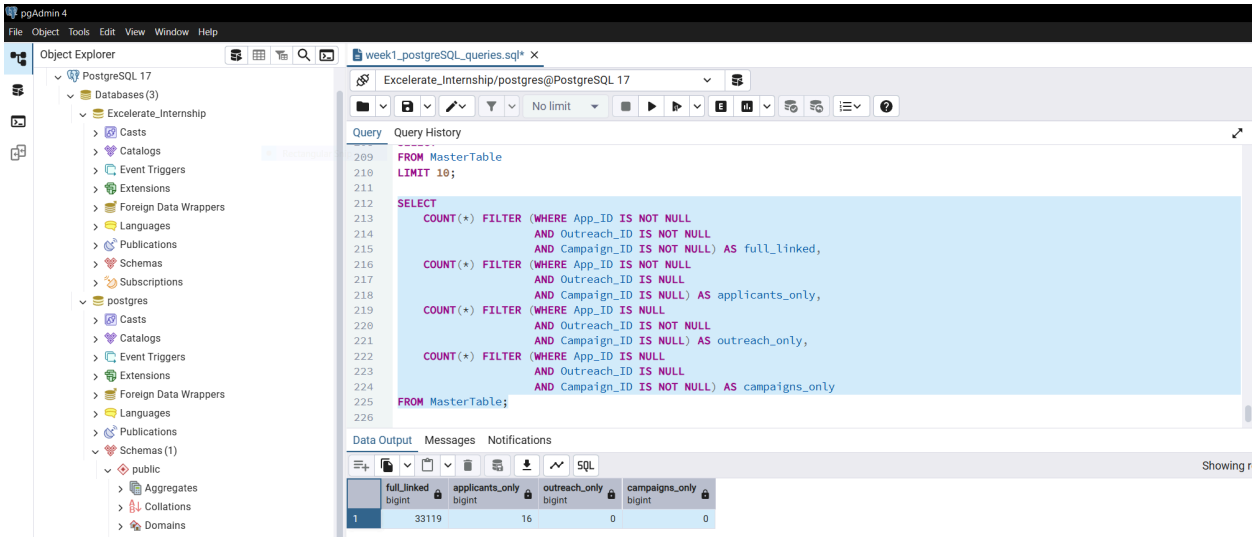


- **Creating Master Table:**









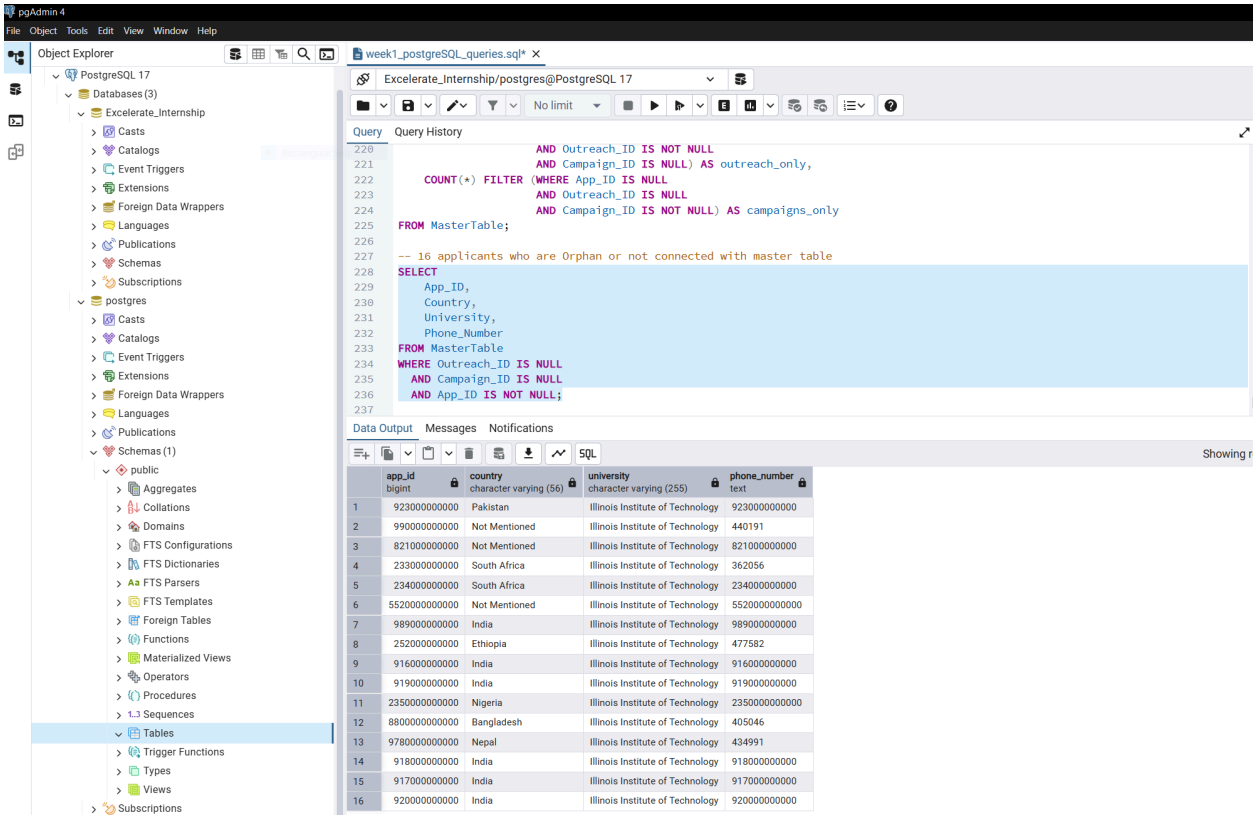
So, here we can see,

1. 33119 under “full_linked” which means rows where we had an App_ID, Outreach_ID, and Campaign_ID all matched.
2. 16 under “applicants_only” depicts 16 applicants exist in our applicant table but have no outreach and no campaign linked.
3. “outreach_only” has 0 which indicates that no outreach rows are floating around without applicant or campaign.

4. “campaigns_only” also has 0 shows, no campaign rows exist without outreach/applicant.

This illustrates that our Master Table (data) is very well connected and reliable where 99.95% of records are fully linked (33,119 out of 33,135). Only 16 applicants are “orphaned”. And these applicants exist in our “CleanApplicantData”, but no one reached out to them and they don’t belong to any campaign.

Furthermore, These are the 16 applicants who could not connect with master table.



Conclusion:

In this project, there are three datasets: CleanApplicantData, Cleaned_CampaignData, and Cleaned_OutreachData. And we cleaned and integrated into a single Master Table using a full outer join. The Master Table provides the full view of applicants, their associated campaigns, and outreach activities, enabling comprehensive analysis of the end-to-end process.

The results of the Master Table showed that out of 33,135 total records, 33,119 were fully connected across applicants, campaigns, and outreach, while only 16 applicants remained disconnected, which is showcasing a high level of data consistency.

The Master Table serves as a single source of truth, simplifies future analysis, ensures data quality, and supports decision-making by offering a holistic view of the applicant engagement lifecycle.