**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.
  + A screenshot of a computer

    AI-generated content may be incorrect.
  + A screenshot of a computer

    AI-generated content may be incorrect.
  + From here we can see, App\_ID has 1632 inconsistent data with 7319 duplicates which is polluting the data integrity. Also 36249 clean data.
  + A screenshot of a computer

    AI-generated content may be incorrect.
  + 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.
  + A screenshot of a computer

    AI-generated content may be incorrect.
* **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.
  + A screenshot of a computer

    AI-generated content may be incorrect.

1. **Data Cleaning:**
   * **Applicant Data:**
     1. 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.
     2. 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.
     3. 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:**
     1. In this dataset, there was a Start\_Date column which we changed the format to Date.
     2. 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:**
     1. “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.
     2. 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.
     3. 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.
2. **PostgreSql Setup Proof:**

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

A screenshot of a computer

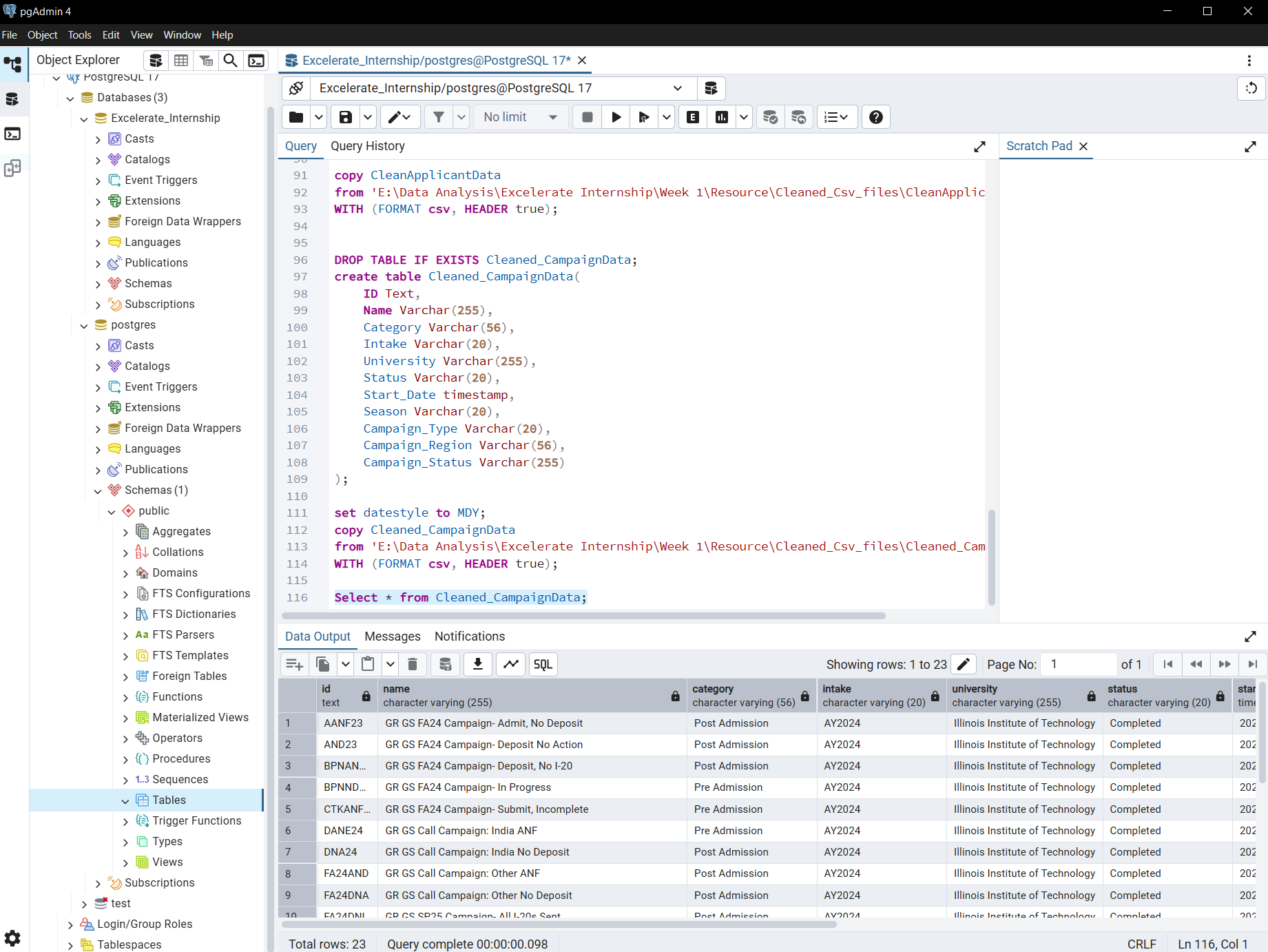
AI-generated content may be incorrect.

View of connected Cleaned Applicant Data.

A screenshot of a computer

AI-generated content may be incorrect.

Also connected the Cleaned\_CampaignData to postgreSql.



Cleaned\_OutreachData is alson ready to find trends and patterns using PostgreSql.A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

* **Exploring the “Clean\_ApplicantData”:**

A screenshot of a computer

AI-generated content may be incorrect.

No inconsistent data is found in “Phone\_Number” column of cleaned applicant data.

A screenshot of a computer

AI-generated content may be incorrect.

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”:**

A screenshot of a computer

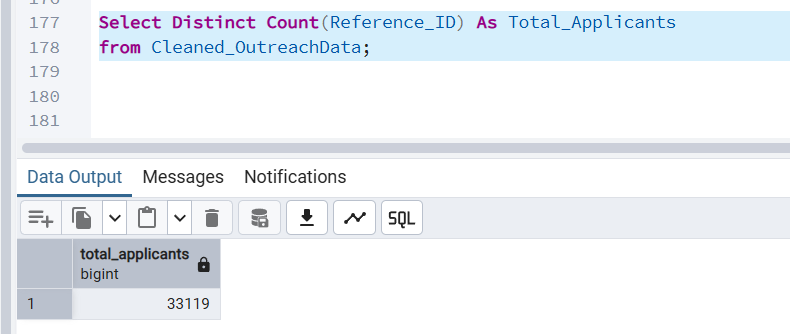
AI-generated content may be incorrect.

These are the callers who has been contacting with applicants.

A screenshot of a computer

AI-generated content may be incorrect.

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”**

A screenshot of a computer

AI-generated content may be incorrect.

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

A screenshot of a computer

AI-generated content may be incorrect.

The primary key of Cleaned\_CampaignData is all unique as we got same row nums.

A screenshot of a computer

AI-generated content may be incorrect.

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

A screenshot of a computer

AI-generated content may be incorrect.

* **Creating Master Table:**

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

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.

A screenshot of a computer

AI-generated content may be incorrect.

**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.