

Internship Report

Data Visualization and Analysis

Week- 2



Date : 16-06-2025

Team Number : 26

Team Members:

1. Khushi Khati
[\(\[khushikhati11@gmail.com\]\(mailto:khushikhati11@gmail.com\)\)](mailto:khushikhati11@gmail.com)
2. Rohit Emmanuel
[\(\[kingrohit2439@gmail.com\]\(mailto:kingrohit2439@gmail.com\)\)](mailto:kingrohit2439@gmail.com)
3. Shomitra Dey
[\(\[soumitradev532@gmail.com\]\(mailto:soumitradev532@gmail.com\)\)](mailto:soumitradev532@gmail.com)

Table of Contents:

Abstract.....	3
1. Introduction.....	4
2. Raw Dataset Exploration.....	5
2.1 Objective.....	5
2.2 Dataset Relationships and Key Identifiers.....	5
2.3 Table: Dataset Overview and Role in Integration.....	6
2.4 Exclusion of the Marketing Dataset.....	6
2.5 Experimental Comparison: Raw vs. Cleaned Master Tables.....	7
3. EDA and Data Visualization.....	8
3.1 Master table created first and then cleaned.....	8
3.2 Master table created using cleaned individual datasets.....	18
4. Master Table Design.....	23
4.1 Objective and Role of the Master Table.....	23
4.2 Selection of the Base Table.....	23
4.3 Dataset Integration Strategy.....	23
4.4 Data Type and Schema Considerations.....	24
4.5 Handling Irrelevant or Faulty Columns.....	24
4.6 Experimental Comparison: Raw vs. Cleaned Versions.....	25
4.7 Summary.....	25
5. Conclusion.....	26
Appendix.....	27
SQL Query Archive.....	27

Abstract

This report, prepared by Team 26, presents the Week 2 progress of our internship project focused on designing and implementing an ETL (Extract, Transform, Load) process. Building upon the exploratory analysis conducted in Week 1, we began by thoroughly examining the structure and relationships across six raw datasets related to learner demographics, cohort timelines, program participation, and user metadata. The primary objective for this week was to integrate these datasets into a unified and consistent Master Table without altering the original raw files. To achieve this, we analyzed key columns, identified relational links such as `learner_id` and `cohort_code`, and documented critical data quality issues, including missing values, inconsistent formats, and orphan records. We also made the deliberate decision to exclude the `marketing_data.csv` dataset from integration due to its lack of relational keys. A major learning point came from conducting an experiment where we built the Master Table twice—once using raw datasets directly, and once using cleaned datasets. This helped us understand the value of preprocessing, as the cleaned version resulted in improved column consistency, fewer NULLs, and greater usability. Throughout this process, we applied structured SQL queries to detect data issues, standardize formats, and validate the integrity of the final Master Table. The completed table now serves as a reliable foundation for future analytical work and dashboards. This report not only highlights our technical decisions but also reflects our iterative learning approach toward building scalable and dependable data pipelines.

Chapter 1

Introduction

The second week of our internship marks a critical transition from exploratory data analysis to practical data integration and transformation. While Week 1 focused on understanding the structure, composition, and quality of individual raw datasets, Week 2 builds upon those insights to design a robust and scalable ETL (Extract, Transform, Load) pipeline that prepares the data for unified analysis.

In this phase, our primary objective is to ensure that the raw datasets—originally collected from multiple independent sources—can be transformed into a consolidated Master Table that adheres to integrity, consistency, and usability standards. This transformation must be done without altering the raw source data, which is preserved for traceability and auditing purposes.

To accomplish this, the report begins with an in-depth exploration of the raw datasets to analyze their structure, detect relational links, and identify potential issues such as missing values, duplicate records, inconsistent data formats, and orphan entries. These issues, if left unaddressed, could have a severe impact on downstream analysis and lead to inaccurate insights.

Following this diagnostic process, we outline a proposed schema for the Master Table—a centralized, cleaned, and relationally sound table that merges relevant attributes from all source datasets. We also draft a detailed ETL procedure, specifying how data should be extracted from the source tables, what transformations are required for cleaning and standardization, and the logic for loading the transformed data into the Master Table.

This week's deliverable culminates with a comprehensive validation step, where the integrity of the Master Table is assessed using data quality checks such as record counts, referential consistency, and format verification. Any discovered anomalies are documented, and the ETL process is refined accordingly to produce a final dataset that is both analytically powerful and trustworthy.

Ultimately, this report serves as a blueprint for data integration in a multi-source environment. It documents our team's approach to problem identification, solution design, and quality assurance—laying the groundwork for more advanced modeling, analytics, and decision-making in the upcoming weeks.

Note: The Marketing Data dataset was excluded from integration during this phase due to its lack of meaningful correlation or relational keys with other datasets.

Chapter 2

Raw Dataset Exploration

2.1 Objective

Understand the structure, relationships, and integration potential of the following datasets:

- user_data.csv
- opp_data.csv
- cohort_data.csv
- marketing_data.csv
- learner_opportunity_raw.csv
- cognito_raw.csv

2.2 Dataset Relationships and Key Identifiers

To establish meaningful relationships between the datasets and construct a coherent Master Table, we initiated a sequence of LEFT JOIN operations using well-defined key columns. The user_dataset served as the base table, as it contained unique learner_id values with no duplicates, making it suitable to act as a primary key for our integration process. This learner_id was first joined with the cognito dataset using the user_id field, ensuring that user profile and authentication metadata could be aligned with demographic information. Next, the user_dataset was joined with learner_opportunity via the same learner_id field to capture enrollment data. This was followed by joining the opportunity table using opportunity_id, and finally linking to the cohort table through the assigned_cohort and cohort_code fields. This layered join architecture allowed us to trace a learner's journey from profile to enrollment to opportunity and cohort timeline, forming a well-connected structure for the Master Table.

Importantly, the marketing_data dataset was excluded from the integration process. Upon analysis, we found that it lacked any relational keys (e.g., learner_id, opportunity_id, or cohort_id) that could link it to other datasets. Since it exists as a standalone dataset with no referential integrity with the rest of the schema, it was deemed unsuitable for inclusion in the ETL pipeline at this stage. Its data may still offer independent insights for future exploratory or campaign-specific analysis but was not integrated into the Master Table due to the absence of logical joins.

2.3 Table: Dataset Overview and Role in Integration

Dataset Name	Key Column(s)	Role in Integration
user_dataset.csv	learner_id	Base table; contains unique learner demographics. Acts as the primary key for joins.
cognito_raw.csv	user_id	Profile and authentication data; joined to user_dataset via learner_id = user_id.
learner_opportunity_raw.csv	enrollment_id, learner_id, assigned_cohort	Tracks learner engagement and cohort assignments.
opp_data.csv	opportunity_id	Program information; connected via opportunity_id from learner_opportunity_raw.
cohort_data.csv	cohort_id, cohort_code	Cohort timelines; joined using assigned_cohort = cohort_code.
marketing_data.csv	(No relational key)	Not included – lacks foreign keys to connect with other datasets.

2.4 Exclusion of the Marketing Dataset

While analyzing potential joins, we determined that marketing_data.csv could not be included in the Master Table design. This dataset contained valuable standalone metrics related to campaign performance, but lacked any keys (such as learner_id, opportunity_id, or cohort_id) to link it with user or engagement data. Therefore, it was excluded from the ETL pipeline, although it may be reused for separate analysis in future weeks.

2.5 Experimental Comparison: Raw vs. Cleaned Master Tables

As part of our learning process, we decided to conduct a comparative experiment:

- First, we followed the instructor's guidance and created the initial Master Table directly from the raw datasets, without cleaning or modifying the data. This helped us simulate a real-world ETL use case where raw data is transformed downstream rather than upfront.
- Then, we cleaned the datasets individually—handling missing values, formatting inconsistencies, and standardizing columns—and created a second Master Table using the cleaned data. This version helped us better understand how upstream cleaning can improve data integrity and reduce post-load transformation complexity.

This dual approach not only validated our ETL logic but also provided a strong practical insight into when and where data cleaning should ideally happen in a pipeline.

The cleaned-data approach proved to be significantly smoother and more reliable. For example, the `cognito_raw` dataset had import issues in the raw form due to column prefix or formatting inconsistencies. This caused multiple fields—including `gender`, `email`, `zip`, `state`, and `birth_date`—to be imported as `NULL`, making them unusable. As a result, these columns were later dropped from the raw version Master Table and then added again after cleaning.

By contrast, in the second version, where datasets were cleaned individually before joining, these issues were avoided entirely. Column names were standardized, formatting issues were resolved during preprocessing, and all useful fields were retained with proper values. This approach not only improved data quality but also reduced the need for post-load fixes, making the ETL pipeline more efficient and easier to manage.

Chapter 3

EDA and Data Visualization

3.1 Master table created first and then cleaned

1) Check Total Rows in the Table

This step confirms the total number of rows in the master table to understand the dataset size.

SQL:

```
SELECT COUNT(*) FROM master_table;
```

	count bigint
1	184710

2) Inspect Table Structure

This step retrieves the column names and data types of the master table to understand its structure.

SQL:

```
SELECT column_name, data_type
```

```
FROM information_schema.columns
```

```
WHERE table_name = 'master_table';
```

	column_name name	data_type character varying			
1	size	integer	14	email	text
2	start_date	date	15	gender	text
3	user_create_date	date	16	city	text
4	user_last_modified_date	date	17	zip	text
5	birth_date	date	18	learner_id	text
6	apply_date	date	19	state	text
7	status	integer	20	country	text
8	end_date	date	21	degree	text
9	opportunity_code	text	22	institution	text
10	tracking_questions	text	23	major	text
11	cohort_code	text	24	opportunity_id	text
12	lo_learner_id	text	25	opportunity_name	text
13	cognito_user_id	text	26	category	text

3) Analyze Category Distribution

This step counts the occurrences of each category in the master table to understand the distribution of opportunity types.

SQL:


```
SELECT category, COUNT(*) AS count
```

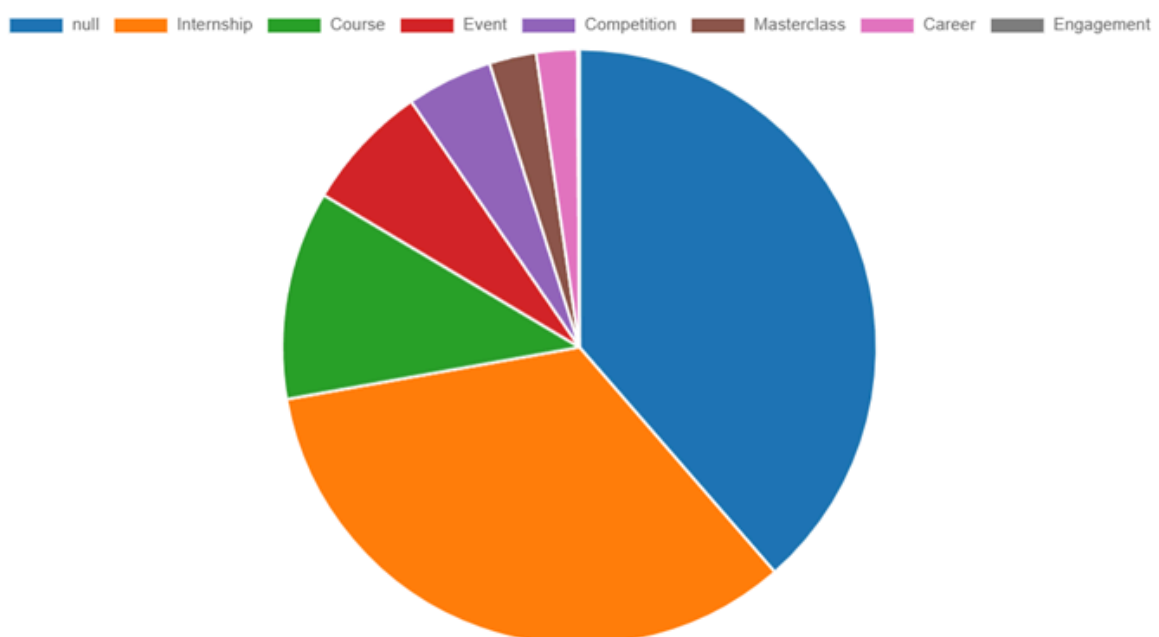
```
FROM master_table
```

```
GROUP BY category
```

```
ORDER BY count DESC;
```

	category text	count bigint
1	[null]	71294
2	Internship	62082
3	Course	20913
4	Event	12745
5	Competition	8642
6	Masterclass	4720
7	Career	4110
8	Engageme...	204

Visualization:



4) Check the Time Range of Opportunities

This step identifies the minimum start_date and maximum end_date in the master table to understand the time period covered by the opportunities.

SQL:

```
SELECT MIN(start_date) AS earliest_start, MAX(end_date) AS latest_end
FROM master_table;
```

	earliest_start date	latest_end date
1	2022-04-15	2026-02-02

5) Analyze Participation by Top 40 Countries

This step counts the number of learners per country in the master_table and limits the results to the top 40 countries based on participant count to identify the most active geographic regions.

SQL:

```
SELECT country, COUNT(*) AS participant_count
```

```
FROM master_table
```

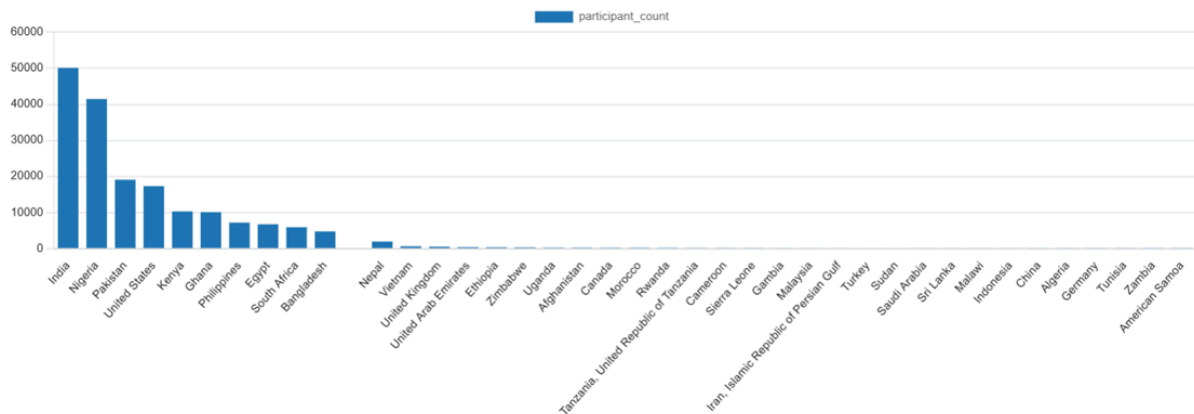
```
GROUP BY country
```

```
ORDER BY participant_count DESC
```

```
LIMIT 40;
```

country text	participant_count bigint		
1 India	50107	21 Morocco	230
2 Nigeria	41480	22 Rwanda	215
3 Pakistan	19088	23 Tanzania, United Republic of Tanza...	185
4 United States	17309	24 Cameroon	184
5 Kenya	10284	25 Sierra Leone	177
6 Ghana	10095	26 Gambia	156
7 Philippines	7185	27 Malaysia	155
8 Egypt	6724	28 Iran, Islamic Republic of Persian Gulf	142
9 South Africa	5899	29 Turkey	136
10 Bangladesh	4736	30 Sudan	130
11 [null]	2275	31 Saudi Arabia	119
12 Nepal	1931	32 Sri Lanka	118
13 Vietnam	613	33 Malawi	114
14 United Kingdom	501	34 Indonesia	114
15 United Arab Emirates	368	35 China	96
16 Ethiopia	328	36 Algeria	95
17 Zimbabwe	303	37 Germany	94
18 Uganda	242	38 Tunisia	86
19 Afghanistan	239	39 Zambia	75
20 Canada	230	40 American Samoa	73

Visualization:



6) Analyze Trends Over Time

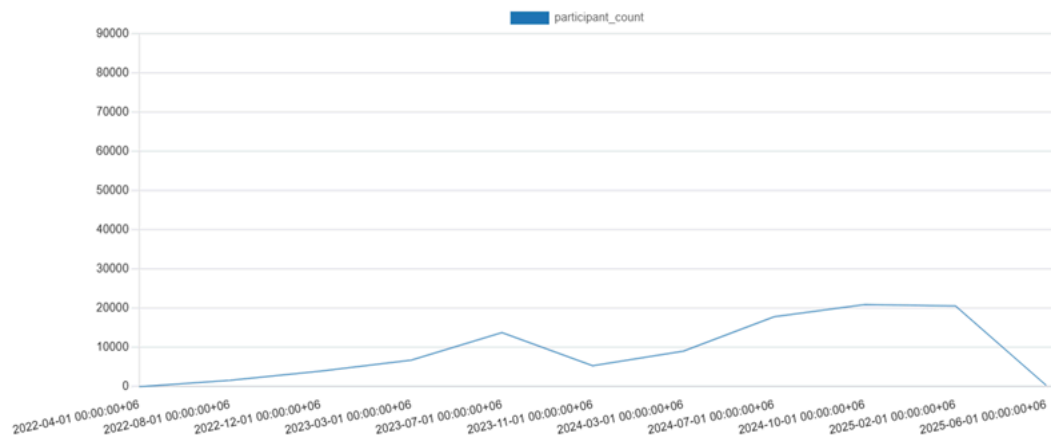
This step groups the data by month from the start_date in the master_table to observe participation trends over time.

SQL:

```
SELECT DATE_TRUNC('month', start_date) AS month, COUNT(*) AS participant_count
FROM master_table
GROUP BY DATE_TRUNC('month', start_date)
ORDER BY month;
```

	month timestamp with time zone	participant_count bigint
1	2022-04-01 00:00:00+06	1
2	2022-08-01 00:00:00+06	1623
3	2022-12-01 00:00:00+06	3977
4	2023-03-01 00:00:00+06	6772
5	2023-07-01 00:00:00+06	13787
6	2023-11-01 00:00:00+06	5337
7	2024-03-01 00:00:00+06	9076
8	2024-07-01 00:00:00+06	17817
9	2024-10-01 00:00:00+06	20949
10	2025-02-01 00:00:00+06	20607
11	2025-06-01 00:00:00+06	338
12	[null]	84426

Visualization:



7) Analyze Participation by Gender

This step counts the number of learners by gender in the master_table to analyze participation distribution across genders.

SQL:

```
SELECT gender, COUNT(*) AS participant_count
```

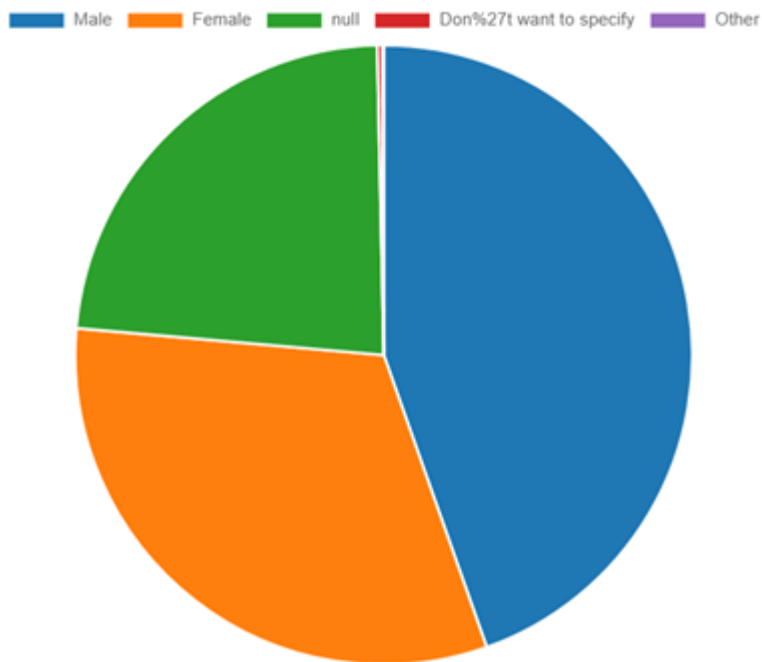
```
FROM master_table
```

```
GROUP BY gender
```

```
ORDER BY participant_count DESC;
```

	gender text	participant_count bigint
1	Male	82348
2	Female	58731
3	[null]	43013
4	Don't want to specify	490
5	Other	128

Visualization:



8) Analyze top 40 Opportunity Size Distribution based on opportunity count

This step summarizes the size column to understand the distribution of opportunity sizes (e.g., number of participants per opportunity) in the master_table.

SQL:

```
SELECT size, COUNT(*) AS opportunity_count
```

```
FROM master_table
```

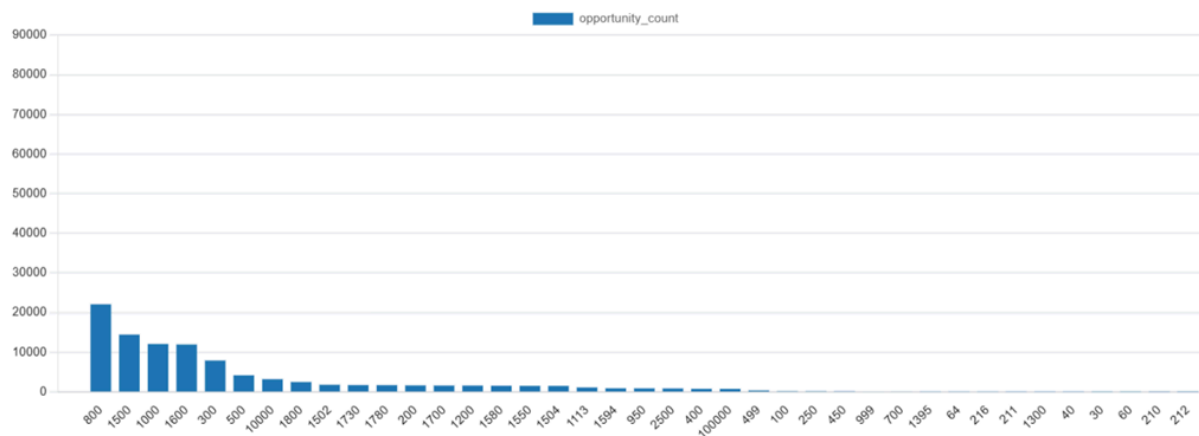
```
GROUP BY size
```

```
ORDER BY opportunity_count desc
```

```
limit 40;
```

	size integer	opportunity_count bigint			
			21	950	885
1	[null]	84426	22	2500	863
2	800	22111	23	400	774
3	1500	14445	24	100000	759
4	1000	12097	25	499	388
5	1600	11974	26	100	227
6	300	7920	27	250	220
7	500	4179	28	450	211
8	10000	3220	29	999	116
9	1800	2495	30	700	105
10	1502	1817	31	1395	61
11	1730	1733	32	64	60
12	1780	1719	33	216	59
13	200	1653	34	211	57
14	1700	1611	35	1300	53
15	1200	1608	36	40	52
16	1580	1564	37	30	44
17	1550	1532	38	60	40
18	1504	1522	39	210	31
19	1113	1112	40	212	29
20	1594	905			

Visualization:



9) Analyze Participation by Degree Level

This step counts the number of learners by degree in the new_master_table to analyze participation across different educational levels.

SQL:

```
SELECT degree, COUNT(*) AS participant_count
```

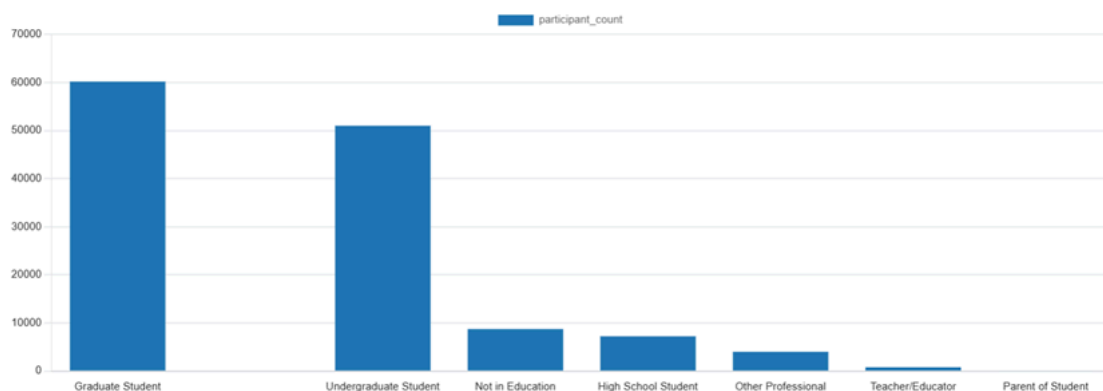
```
FROM new_master_table
```

```
GROUP BY degree
```

```
ORDER BY participant_count DESC;
```

	degree text	participant_count bigint
1	Graduate Student	60190
2	[null]	52707
3	Undergraduate Student	51008
4	Not in Education	8730
5	High School Student	7218
6	Other Professional	3986
7	Teacher/Educator	772
8	Parent of Student	99

Visualization:



10) Analyze Participation by Major

This step counts the number of learners by major in the new_master_table to analyze participation across different academic or professional fields.

SQL:

```
SELECT major, COUNT(*) AS participant_count
```

```
FROM new_master_table
```

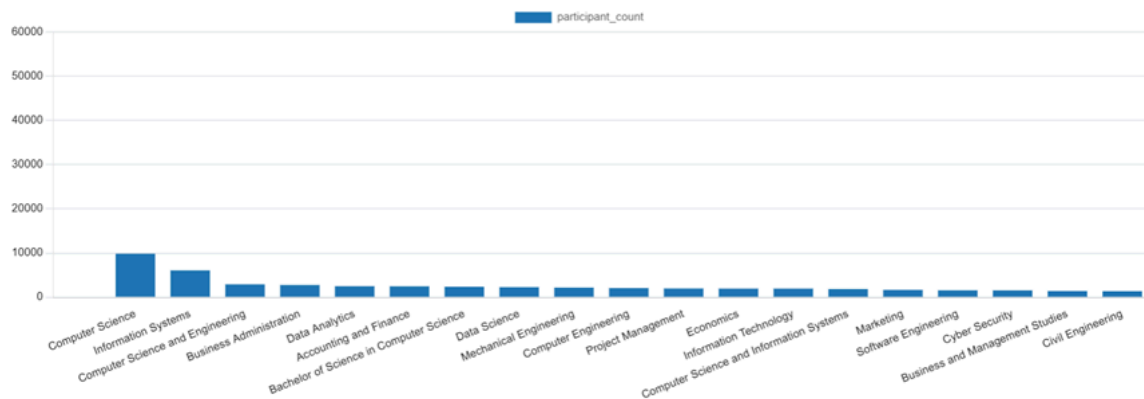
```
GROUP BY major
```

```
ORDER BY participant_count DESC
```

```
LIMIT 20;
```

	major text	participant_count bigint
1	[null]	52710
2	Computer Science	9842
3	Information Systems	6080
4	Computer Science and Engineering	2913
5	Business Administration	2770
6	Data Analytics	2492
7	Accounting and Finance	2470
8	Bachelor of Science in Computer Science	2376
9	Data Science	2274
10	Mechanical Engineering	2170
11	Computer Engineering	2065
12	Project Management	1979
13	Economics	1951
14	Information Technology	1947
15	Computer Science and Information Systems	1856
16	Marketing	1654
17	Software Engineering	1561
18	Cyber Security	1548
19	Business and Management Studies	1394
20	Civil Engineering	1357

Visualization:



11) Analyze Participation by Institution

This step counts the number of learners by institution in the master_table to analyze participation across different institutions.

SQL:

```
SELECT institution, COUNT(*) AS participant_count
```

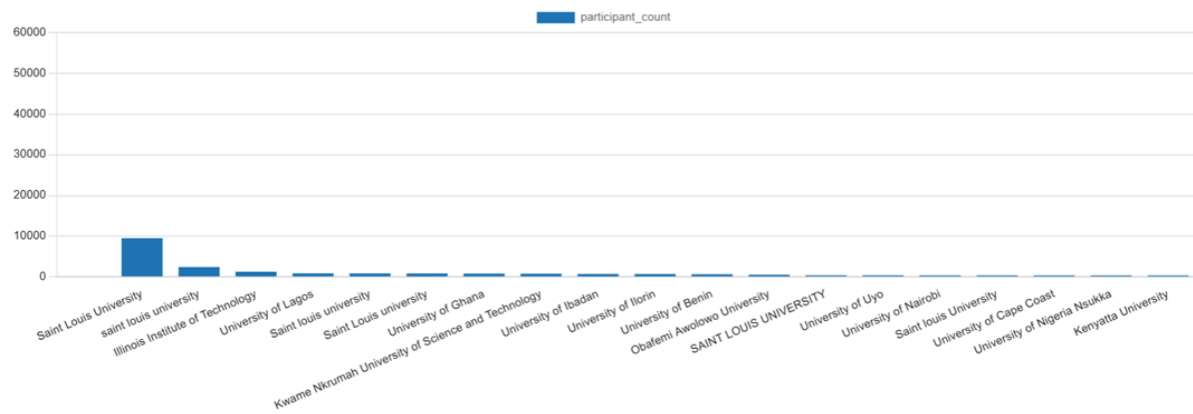
```
FROM new_master_table
```

```
GROUP BY institution
```

```
ORDER BY participant_count DESC
```

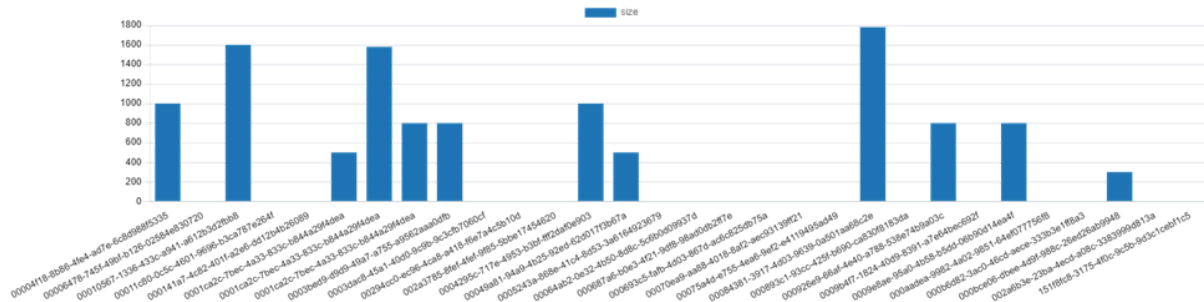
```
LIMIT 20;
```


Visualization:



3.2 Master table created using cleaned individual datasets


```
SELECT * FROM Master_table LIMIT 30 (X axis learner_id, Y axis size)
```



1) Total Rows in the Table

SQL:

```
SELECT COUNT(*) FROM Master_table;
```

	count bigint 
1	184710

SQL:

```
SELECT column_name, data_type
```

FROM information_schema.columns

WHERE table_name = 'Master_table';

2) Analyze Category Distribution

This step counts the occurrences of each category in the master table to understand the distribution of

opportunity types.

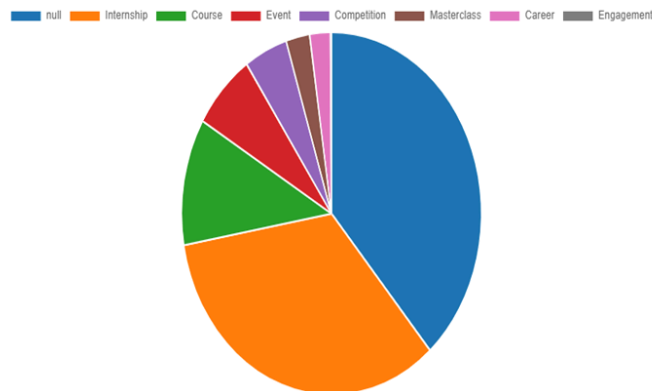
SQL:

SELECT category, COUNT(*) AS count Visualization:

FROM master table

GROUP BY category

ORDER BY count DESC



3) Check Time Range of Opportunities

This step identifies the minimum start_date and maximum end_date in the master_table to understand

the time period covered by the opportunities.

SQL:

```
SELECT MIN(start_date) AS earliest_start, MAX(end_date) AS latest_end  
FROM master_table;
```

	earliest_start date	latest_end date
1	2022-06-09	2026-03-06

4) Analyze Participation by Top 40 Countries

This step counts the number of learners per country in the master_table and limits the results to the top

40 countries based on participant count to identify the most active geographic regions.

SQL:

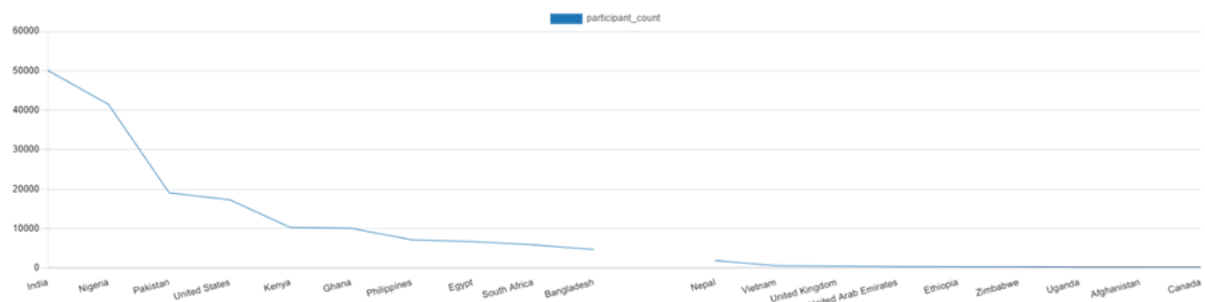
```
SELECT country, COUNT(*) AS participant_count  
FROM master_table
```

GROUP BY country

ORDER BY participant_count DESC

LIMIT 20;

	country text	participant_count bigint
1	India	50107
2	Nigeria	41480
3	Pakistan	19088
4	United States	17309
5	Kenya	10284
6	Ghana	10095
7	Philippines	7185
8	Egypt	6724
9	South Africa	5899
10	Bangladesh	4736
11	[null]	2275
12	Nepal	1931
13	Vietnam	613
14	United Kingdom	501
15	United Arab Emirates	368
16	Ethiopia	328
17	Zimbabwe	303
18	Uganda	242
19	Afghanistan	239
20	Canada	238



5) Analyze Trends Over Time

This step groups the data by month from the start_date in the master_table to observe participation

trends over time.

SQL:

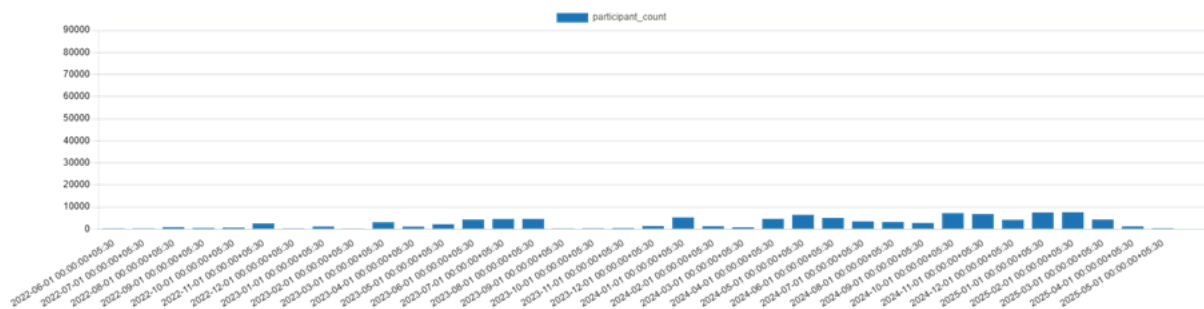
```
SELECT DATE_TRUNC('month', start_date) AS month, COUNT(*) AS participant_count
```

```
FROM master_table
```

```
GROUP BY DATE_TRUNC('month', start_date)
```

```
ORDER BY month;
```

	month timestamp with time zone	participant_count bigint
1	2022-06-01 00:00:00+05:30	2
2	2022-07-01 00:00:00+05:30	285
3	2022-08-01 00:00:00+05:30	863
4	2022-09-01 00:00:00+05:30	474
5	2022-10-01 00:00:00+05:30	644
6	2022-11-01 00:00:00+05:30	2560
7	2022-12-01 00:00:00+05:30	21
8	2023-01-01 00:00:00+05:30	1114
9	2023-02-01 00:00:00+05:30	222
10	2023-03-01 00:00:00+05:30	3106
11	2023-04-01 00:00:00+05:30	1078
12	2023-05-01 00:00:00+05:30	2165
13	2023-06-01 00:00:00+05:30	4324
14	2023-07-01 00:00:00+05:30	4551
15	2023-08-01 00:00:00+05:30	4575



6) Analyze Participation by Gender

This step counts the number of learners by gender in the master table to analyze participation

distribution across genders.

SQL:

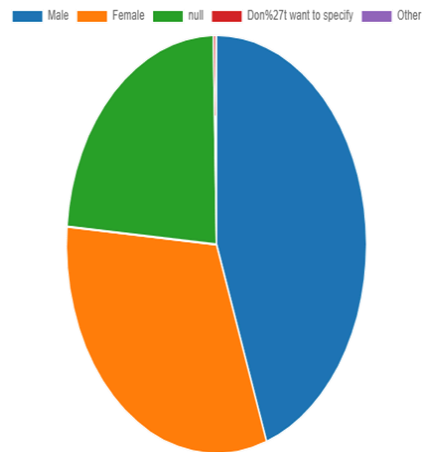
```
SELECT gender, COUNT(*) AS participant_count
```

```
FROM master_table
```

```
GROUP BY gender
```

```
ORDER BY participant_count DESC;
```

	gender text	participant_count bigint
1	Male	82348
2	Female	58731
3	[null]	43013
4	Don't want to specify	490
5	Other	128



Chapter 4

Master Table Design

This chapter describes the structure, logic, and decision-making involved in designing the Master Table, which serves as the consolidated foundation for downstream analysis and reporting.

4.1 Objective and Role of the Master Table

The Master Table was developed to represent a unified, structured view of all learner-related information drawn from multiple raw datasets. It combines demographic details, program participation, cohort timelines, and enrollment history in a single, analysis-ready format. This central dataset plays a pivotal role in the ETL pipeline, acting as the final cleaned output after data extraction and transformation.

4.2 Selection of the Base Table

We selected the `user_dataset` as the base of the Master Table due to the reliability of its `learner_id` field, which contained unique, non-null, and duplicate-free values. This made `learner_id` a strong candidate for a primary key, ensuring one record per learner and enabling consistent joins across other datasets.

4.3 Dataset Integration Strategy

The Master Table was constructed by progressively joining relevant datasets to the base table (`user_dataset`) using LEFT JOIN operations:

- From the `user_dataset`, we retained fields like `country`, `degree`, `institution`, and `major` to describe learner demographics and background.
- From the `cognito_raw` dataset, we initially attempted to include fields such as `email`, `gender`, `birth_date`, and `city`. However, due to formatting issues, many of these columns were populated entirely with NULL values and were later dropped during transformation. However, they were re-added after the Master table was cleaned.
- From the `learner_opportunity_raw` dataset, we extracted important fields like `apply_date`, `status`, and `assigned_cohort`, which help capture application timelines and program engagement.
- From the `opp_data` dataset, fields like `opportunity_name`, `category`, and other program descriptors were added to provide context to the learner's application.
- From the `cohort_data` dataset, we brought in cohort-level details such as `start_date`, `end_date`, and `size`, helping link learners to specific learning timelines.

4.4 Data Type and Schema Considerations

We ensured data type consistency across all fields. Categorical columns (e.g., country, degree, category) were cast to TEXT or VARCHAR types. All dates were standardized to TIMESTAMP format to maintain compatibility across joins. Numerical fields, such as status and cohort size, were kept as INTEGER, depending on range and usage.

This consistency was critical to prevent transformation errors and ensure smooth data validation later in the ETL process.

4.5 Handling Irrelevant or Faulty Columns

As detailed earlier, several fields from the cognito_raw dataset were dropped from the Master Table due to prefix or header format issues during import. These columns included:

- email, gender, birth_date, zip, state, city
- user_create_date, user_last_modified_date

These fields were entirely NULL and had no analytical value in their corrupted state. Their removal helped reduce clutter and improve the table's quality.

SQL:

```
ALTER TABLE master_table
DROP COLUMN cognito_user_id,
DROP COLUMN email,
DROP COLUMN gender,
DROP COLUMN user_create_date,
DROP COLUMN user_last_modified_date,
DROP COLUMN birth_date,
DROP COLUMN city,
DROP COLUMN zip,
DROP COLUMN state;
```


4.6 Experimental Comparison: Raw vs. Cleaned Versions

To deepen our understanding of ETL design, we conducted an experiment where we created the Master Table twice:

1. Raw Master Table: Created by directly joining uncleaned datasets, as per the initial instructor's instructions.
2. Cleaned Master Table: Built using pre-cleaned datasets, where missing values, formatting issues, and inconsistent casing were addressed beforehand.

The cleaned version proved to be far more consistent and easier to work with, especially since it avoided the NULL import problems we faced in the `cognito_raw` dataset. This version also preserved valuable columns that were otherwise unusable in the raw pipeline. Through this comparison, we learned that cleaning data prior to integration can significantly improve pipeline reliability and minimize post-processing work.

4.7 Summary

The Master Table design emphasizes data integrity, relevance, and usability. By using a robust key (`learner_id`), choosing compatible fields, and removing broken columns, we built a unified structure that supports both analysis and future transformation needs. It now serves as the primary foundation for validation, visualization, and insight extraction via dashboard creation in the upcoming stages of this project.

Chapter 5

Conclusion

In Week 2, our team made significant progress in transitioning from exploratory analysis to practical data integration. We began by revisiting all raw datasets to understand their structure, identify relationships, and document data quality issues. Through this process, we established that `learner_id` in the `user_dataset` was the most reliable unique identifier and used it as the anchor for building our Master Table.

We encountered several challenges, including missing values, inconsistent data formats, and import issues—particularly in the `cognito_raw` dataset, where key columns were entirely NULL due to prefix-related formatting errors. Additionally, we determined that the `marketing_data` dataset could not be included in the Master Table due to its lack of relational keys and meaningful joins.

To deepen our understanding, we experimented by creating two versions of the Master Table: one using raw datasets directly (as per instructions) and another using pre-cleaned datasets. The comparison clearly showed that cleaning datasets prior to integration yields more accurate, complete, and usable results—an important lesson for real-world ETL workflows.

The finalized Master Table integrates learner demographics, program applications, opportunity details, and cohort timelines. It is structured with consistent data types and validated for duplicates, missing values, and referential integrity. This table now stands as the core output of our ETL process and will serve as the foundation for future reporting, visualization, and advanced analytics in the upcoming weeks.

Our work this week has laid the groundwork for a robust data pipeline, emphasizing the importance of thoughtful design, iterative validation, and the value of clean, connected data.

Appendix

SQL Query Archive

As part of the ETL and Master Table construction process, numerous SQL queries were written to explore, validate, and transform the datasets. These queries cover tasks such as data inspection, missing value detection, duplicate checks, JOIN operations, and column cleaning.

To maintain clarity and reusability, all SQL scripts have been compiled into organized files and uploaded to a shared folder.

Access the complete SQL query archive here:

[SQL Queries](#)

{Team Name: FourSight} Team
Charter

Team Members	<p>1. Khushi Khati (khushikhati11@gmail.com)</p> <p>2. Rohit Emmanuel (kingrohit2439@gmail.com)</p> <p>3. Shomitra Dey (Soumitradev532@gmail.com)</p>
Team Lead	<p>Khushi Khati (khushikhati11@gmail.com)</p>

<p>Team Members Roles and Responsibilities</p>	<ol style="list-style-type: none"> <p>1. Khushi Khati (khushikhati11@gmail.com)</p> <p>Team Lead: Represents the team to sponsor, via email and on calls, to minimize communication errors.</p> <p>2. Rohit Emmanuel (kingrohit2439@gmail.com)</p> <p>Project Scribe: Responsible for taking meeting minutes and distributing notes/assignments. Can assist the Team Lead in drafting emails and communication between the sponsor and the group.</p> <p>3. Shomitra Dey (Soumitradev532@gmail.com)</p> <p>Project Lead: Responsible for holding the group accountable for meeting deadlines and ensuring that the project deliverables are being met.</p>
---	--

<p>Mission, Vision Objectives & Core Values</p>	<p><u>Mission:</u> To deliver an impressive and valuable impact by focusing on innovation and problem-solving. Through a well-structured project plan, we aim to provide tangible outcomes, primarily a comprehensive report, that can be effectively implemented. Close collaboration with colleagues throughout the process is central to our mission, ensuring our work remains aligned with their goals.</p> <p><u>Vision:</u> Our vision of success lies in fostering a collaborative and inclusive team environment built on open communication and equal contribution. Despite language and cultural differences, we are committed to working together with strength and efficiency. Beyond just completing deliverables, we view this project as an opportunity to build long-term professional skills and meaningful connections. Success for us means achieving both project excellence and growth through networking and teamwork.</p>
--	---

	<p><u>Core Values:</u> Honesty, Ownership, Dedication, Mutual Respect, Creative Thinking, Collaboration, Growth Mindset, Transparency</p>
--	--

Internal Checks, Balances, and Reviews	<p>One of our team's core strengths lies in our consistent willingness to support one another. We value and respect each member's input, make decisions collectively, and ensure that everyone feels included and appreciated. These qualities foster a collaborative environment grounded in mutual respect and shared responsibility. Our strengths naturally complement each other, allowing us to engage in professional and efficient discussions without conflict. Notably, when a team member is unavailable or overwhelmed, others step in proactively and selflessly to maintain momentum. This proactive attitude, combined with a strong work ethic and mutual accountability, sets our team apart in terms of</p>
	<p>reliability and cohesion.</p>

<p>Operations:</p> <ul style="list-style-type: none"> • Assignments • Meetings • Communication Guidelines • Status Updates • Deadlines 	<p>Team Meetings.</p> <ul style="list-style-type: none"> • Weekly Meetings: Tuesdays, Saturdays, and Sundays at 8:00 PM IST on Google Meet. Purpose: Progress updates, addressing blockers, tracking deliverables, and problem-solving. <p>Meeting Structure.</p> <ul style="list-style-type: none"> • Agenda (shared 24 hours in advance). • Notes/minutes recorded and shared (within 24 hours). • Action items with owners and due dates. <p>Sub-Team Responsibilities.</p> <p>Each sub-team or individual is expected to:</p> <ul style="list-style-type: none"> • Proactively communicate status updates or blockers. • Document all work in a shared workspace (Google Docs). • Complete assigned tasks by the set deadlines or communicate delays in advance.
---	---