WEEK 2: Data Quality Report

TEAM NAME: 0704 DVA | TEAM 06

Team Members

Name	Email ID
Ushna Adnan	Ushna859@gmail.com
Muna Saha	Munasaha369@gmail.com
Divya P	Divyadp0285@gmail.com
Salman Rahobar	Kazisalmanrahobar@gmail.com
Ritik Mathpal	Work.ritikmathpal@gmail.com
Shoaib Salman	Shoaibsalman28@gmail.com
Syed Rayyan	Syedramsey18@gmail.com
Mary Olurinola	Olurinolamary1@gmail.com

Table of Contents

Chapter 1: Introduction

1. 1. Purpose of the Report

In today's data-driven landscape, ensuring the accuracy, completeness, and consistency of data is crucial for meaningful insights and decision-making. The Data Quality Report serves as a comprehensive documentation of the data preparation process, highlighting the steps taken to clean, validate, and transform raw datasets into a structured format suitable for analysis and visualization.

This report outlines the key challenges encountered in the data, the methodologies applied to address inconsistencies, and the structured approach used to build a Master Table that integrates multiple datasets efficiently.

1. 2. Scope of the Report

The scope of this report covers six datasets that were provided for analysis:

- **Learner/User Data** – Contains information about individual learners, including their education background, country, and institution.
- **Opportunity Data** – Details various learning opportunities, including opportunity names, categories, and codes.
- **Cohort Data** – Tracks cohort-based learning programs with details such as cohort codes, start and end dates, and cohort sizes.
- **Marketing Data** – Captures advertising performance metrics, including campaign reach, engagement, and costs.
- **Learner Opportunity Data** – Maps learners to specific opportunities, tracking their application status and cohort assignments.
- **Cognito Data** – Includes authentication and profile metadata such as email, gender, location, and account activity timestamps.

Each dataset undergoes exploratory data analysis (EDA) to identify missing values, duplicates, inconsistencies, and outliers. This is followed by data cleaning and transformation, ensuring that only high-quality data is used for analysis.

1. 3. Objectives of Data Cleaning & Transformation

The primary objectives of this report are:

1. **Assess Data Quality Issues** – Identify missing values, duplicate records, incorrect formats, and anomalies.
2. **Standardize & Transform Data** – Apply rules to clean, format, and structure data consistently.
3. C**reate a Master Table** – Integrate data from multiple sources to establish relationships for analysis.
4. **Ensure Data Accuracy & Integrity** – Validate records, remove inconsistencies, and maintain referential integrity.
5. P**repare Data for Analysis & Dashboarding** – Provide a structured, highquality dataset that can be used for visualization and reporting.

Chapter 2: Data Overview

2. 1. Learner Dataset:

1. Source

Learner data comes from the platform’s user registration and profile system, capturing details about learners' education, country, and academic background.

2. Key Fields

- **learner_id (Primary Key)** – Unique identifier for each learner.
- **country** – Country of residence.
- **degree** – Level of education (e.g., Bachelor's, Master's, Diploma).
- **institution** – Name of the educational institution.
- **major** – The field of study of the learner.

3. Relationships

- PK: learner_id
- FK: Linked to Learner Opportunity Data via learner_id.

4. Initial Data Summary

- Total Rows: 129259
- Total Columns: 5

	total_rows bigint
1	129259

	column_name name
1	learner_id
2	country
3	degree
4	institution
5	major

5. Data Types

	column_name name	data_type character varying
1	learner_id	uuid
2	country	character varying
3	degree	text
4	institution	text
5	major	text

6. Issues Identified:

- Missing Values: Yes, found in institution and major.
- Duplicates: No duplicates in learner_id.
- Outliers: Some unusually long text in institution and major.

2. 2. Opportunity Dataset:

1. Source

This dataset contains structured data about various learning opportunities available in the system, including courses, workshops, and events.

2. Key Fields

- **opportunity_id (Primary Key) –** Unique identifier for each opportunity.
- **opportunity_name –** Title of the opportunity.
- **category –** Type of opportunity (e.g., event, course, challenge).
- **opportunity_code –** Internal reference code for tracking.

3. Relationships

- PK: opportunity_id
- FK: Linked to Learner Opportunity Data via opportunity_id.

4. Initial Data Summary

- Total Rows: 187
- Total Columns: 5

	total_rows bigint
1	187

	column_name name
1	learner_id
2	country
3	degree
4	institution
5	major

5. Data Types

	column_name name	data_type character varying
1	opportunity_id	character varying
2	opportunity_name	text
3	category	character varying
4	opportunity_code	character varying
5	tracking_questions	jsonb

6. Issues Identified:

- Missing Values: None.
- Duplicates: No duplicates in opportunity_id, but some repeated opportunity names.
- Outliers: No major outliers found.

2. 3. Cohort Dataset:

1. Source

Tracks cohort-based learning programs, grouping learners into structured learning paths.

2. Key Fields

- **cohort_code (Primary Key) –** Unique identifier for each cohort.
- **start_date –** The timestamp when the cohort begins**.**
- **end_date –** The timestamp when the cohort ends.
- **size –** The number of learners in the cohort.

3. Relationships

- PK: cohort_code
- FK: Linked to Learner Opportunity Data via assigned_cohort.

4. Initial Data Summary

- Total Rows: 639
- Total Columns: 4

	total_rows bigint
1	639

	column_name name
1	cohort_code
2	start_date
3	end_date
4	size

5. Data Types

	column_name name	data_type character varying
1	cohort_code	character varying
2	start_date	timestamp without time zone
3	end_date	timestamp without time zone
4	size	integer

6. Issues Identified:

- Missing Values: None.
- Duplicates: No duplicate cohort codes.
- Outliers: Found in size (some extremely large values).

2. 4. Marketing Dataset:

1. Source

This dataset contains advertising campaign performance data, measuring reach, engagement, and costs.

2. Key Fields

- **campaign_name –** Name of the marketing campaign.
- **delivery_status –** Status of the campaign (Active, Completed).
- **reporting_starts / reporting_ends –** Start and end dates of campaign tracking.
- **sum_of_reach –** Number of unique users who saw the ad.
- **sum_of_amount_spent –** Budget spent on the campaign (AED).

3. Relationships

- This dataset is independent and does not directly link to other datasets.

4. Initial Data Summary

5. Data Types

- Total Rows: 89
- Total Columns: 13

	total_rows bigint
1	89

	column_name name
1	delivery_status
2	ad_account_name
3	delivery_level
4	result_type
5	reporting_starts
6	reporting_ends
7	count_of_results
8	sum_of_reach
9	sum_of_outbound_clicks
10	sum_of_landing_page_views
11	sum_of_cost_per_result
12	sum_of_cpc
13	sum_of_amount_spent

	column_name name	data_type character varying
1	delivery_status	character varying
2	ad_account_name	character varying
3	delivery_level	character varying
4	result_type	character varying
5	reporting_starts	date
6	reporting_ends	date
7	count_of_results	integer
8	sum_of_reach	integer
9	sum_of_outbound_clicks	integer
10	sum_of_landing_page_views	integer
11	sum_of_cost_per_result	numeric
12	sum_of_cpc	numeric
13	sum_of_amount_spent	numeric

6. Issues Identified:

- Missing Values: reporting_ends contained placeholders (#######), replaced with NULL.
- Duplicates: No duplicate campaigns.
- Outliers: High spending values observed in sum_of_amount_spent.

2. 5. Learner Opportunity Dataset:

1. Source

Tracks which learners have enrolled in specific opportunities, acting as a bridge between Learner Data and Opportunity Data.

2. Key Fields

- **enrollment_id (Primary Key) –** Unique identifier for each enrollment.
- **learner_id (Foreign Key) –** Learner linked to the opportunity.
- **opportunity_id (Foreign Key) –** Opportunity the learner enrolled in.
- **assigned_cohort –** Cohort assignment for the learner.
- **apply_date –** Timestamp of application.

3. Relationships

- PK: enrollment_id
- FKs: Links to learner_id, opportunity_id, and assigned_cohort.

5. Data Types

4. Initial Data Summary

- Total Rows: 113602
- Total Columns: 5

	total_rows bigint
1	113602

	column_name name
1	enrollment_id
2	learner_id
3	assigned_cohort
4	apply_date
5	status

	column_name name	data_type character varying
1	enrollment_id	uuid
2	learner_id	character varying
3	assigned_cohort	character varying
4	apply_date	timestamp without time zone
5	status	integer

6. Issues Identified:

- Missing Values: assigned_cohort had some NULL values.
- Duplicates: No duplicate enrollment_id.
- Outliers: No major outliers found.

2. 6. Cognito Dataset:

1. Source

This dataset contains authentication and user profile metadata.

2. Key Fields

- **user_id (Primary Key) –** Unique identifier for each user.
- **email –** User's email address.
- **gender –** User gender (Male/Female/NULL).
- **usercreatedate –** Account creation timestamp.
- **birthdate –** User's date of birth.

3. Relationships

- PK: user_id
- FK: Could be linked to Learner Data via email.

4. Initial Data Summary

- Total Rows: 129178

5. Data Types

- Total Columns: 9

	total_rows bigint
1	129178

	column_name name
1	user_id
2	email
3	gender
4	usercreatedate
5	userlastmodifieddate
6	birthdate
7	city
8	zip
9	states

	column_name name	data_type character varying
1	user_id	uuid
2	email	character varying
3	gender	character varying
4	usercreatedate	timestamp without time zone
5	userlastmodifieddate	timestamp without time zone
6	birthdate	date
7	city	character varying
8	zip	text
9	states	character varying

6. Issues Identified:

- Missing Values: birthdate, city, and states contained NULL values.
- Duplicates: No duplicate user_id.
- Outliers: Some incorrect date values were replaced.

Chapter 3: Data Validation & Transformation

3. 1. Learner Dataset:

```
--QUERY TO VIEW THE NULL VALUES OF EACH COLUMN
SELECT
        SUM(CASE WHEN learner_id IS NULL THEN 1 ELSE 0 END) AS
        missing_learner_id,
    SUM(CASE WHEN country IS NULL THEN 1 ELSE 0 END) AS missing_country,
        SUM(CASE WHEN degree IS NULL THEN 1 ELSE 0 END) AS
missing_degree,     SUM(CASE WHEN institution IS NULL THEN 1 ELSE 0 END)
AS
        missing_institution,
        SUM(CASE WHEN major IS NULL THEN 1 ELSE 0 END) AS missing_major
FROM learner_data;
```

	missing_learner_id bigint	missing_country bigint	missing_degree bigint	missing_institution bigint	missing_major bigint
1	0	2275	52693	52693	52694

```
--SETTING ALL THE NULL VALUES WITH UNKNOWN
UPDATE learner_data
SET country = 'Unknown'
WHERE country IS NULL;
```

5. Data Types

```
UPDATE learner_data
SET degree = 'Unknown'
WHERE degree IS NULL;
UPDATE learner_data
SET institution = 'Unknown'
WHERE institution IS NULL;
UPDATE learner_data
SET major = 'Unknown'
WHERE major IS NULL;
```

--HANDLING THE COUNTRY COLUMN

--Query to view country names with comma (,)

```
SELECT DISTINCT country
FROM learner_data
WHERE country LIKE '%,%';
```

	country character varying (50)
1	Virgin Islands, British
2	Congo, The Democratic Republic of the Co...
3	Bolivia, Plurinational State of
4	Venezuela, Bolivarian Republic of Venezuela
5	Iran, Islamic Republic of Persian Gulf
6	Tanzania, United Republic of Tanzania
7	Palestinian Territory, Occupied
8	Korea, Republic of South Korea
9	Virgin Islands, U.S.

```
UPDATE learner_data
SET country = SPLIT_PART(country, ',', 1)
WHERE country LIKE '%,%';

UPDATE learner_data
SET country = 'USA'
WHERE country IN ('United States of America', 'United States', 'U.S.', 'America');

UPDATE learner_data
SET country = 'UK'
WHERE country IN ('United Kingdom', 'England', 'Great Britain');

UPDATE learner_data
SET country = 'South Korea'
```

```
--
WHERE country IN ('Korea, Republic of', 'Korea Republic', 'South Korea');
UPDATE learner_data
SET country = 'Iran'
WHERE country IN ('Iran, Islamic Republic of', 'Iran Persian Gulf');
UPDATE learner_data
SET country = 'Côte d''Ivoire'
WHERE country = 'Cote d%27Ivoire';
```

```
--HANDLING THE DEGREE COLUMN
--Query to View All Unique Degree Names
SELECT DISTINCT degree FROM learner_data ORDER BY degree;
```

	degree text
1	Graduate Student
2	High School Student
3	Not in Education
4	Other Professional
5	Parent of Student
6	Teacher/Educator
7	Undergraduate Student
8	Unknown

```
-- Standardize Undergraduate/Bachelor's Degree
UPDATE learner_data
SET degree = 'Bachelor'
WHERE degree = 'Undergraduate Student';
-- Standardize Graduate Degree (Master's or PhD)
UPDATE learner_data
SET degree = 'Graduate'
WHERE degree = 'Graduate Student'
```

Standardize High School

```
UPDATE learner_data
SET degree = 'High School'
WHERE degree = 'High School Student';
```

```
-- Group "Not in Education" and other similar values
UPDATE learner_data
SET degree = 'Other'
WHERE degree IN ('Not in Education', 'Other Professional', 'Parent of Student', 'Teacher/Educator');
```

```
--To see if all degrees look good
SELECT degree, COUNT(*) FROM learner_data GROUP BY degree ORDER BY COUNT(*) DESC;
```

	degree text	count bigint
1	Unknown	52693
2	Graduate	31806
3	Bachelor	30709
4	Other	9942
5	High School	4109

```
--HANDLING INSTITUTION COLUMN
--Find the most frequent institutions
SELECT institution, COUNT(*)
FROM learner_data
GROUP BY institution
ORDER BY COUNT(*) DESC
LIMIT 10;
```

--

	institution text	count bigint
1	Unknown	52694
2	Saint Louis University	2163
3	University of Lagos	605
4	Illinois Institute of Technology	553
5	University of Ghana	524
6	saint louis university	498
7	University of Ibadan	461
8	University of Ilorin	425
9	Kwame Nkrumah University of Science and Technology	410
10	University of Benin	400

--Query to check distint institution

```
SELECT DISTINCT institution FROM learner_data ORDER BY institution;
```

	institution text
1	-
2	----
3	%22i Have Completed My Intermediate From Govt. Ghulam Hussain Hidayatullah Higher Secondary School, Hyderabad
4	(Amu) Alexanderia Of Midea Institute
5	(Euro Mediterranean University
6	.
7	..
8	...
9	@Pir Abdul Qadir Shah Jilani Institute Of Medical Science Gambat
10	+2 B M High School Pinjrawan
11	02 Academy
12	1
13	1. Hussaini Adamu Federal Polytechnic Kazaure. 2. Collage Of Health Sciences And Technology Tsafe.
14	10
15	100
16	10alytics Academy, Uk
17	10th Ramadan University

Query to Identify Placeholder Values

```
SELECT DISTINCT institution

FROM learner_data

WHERE institution IN ('-', '----', '..', '...', '.', '.....');
```

--Query to Identify Numeric Values

```
SELECT DISTINCT institution
FROM learner_data
WHERE institution ~ '^[0-9]+$';
```

	institution text
1	-
2	----
3	.
4	..
5	...

	institution text
1	1
2	10
3	100
4	12
5	123
6	12344
7	2
8	2007
9	2010
10	2015
11	2016
12	234234
13	32
14	60
15	8989809
16	9

--

Query to Identify Institutions with Encoded Characters

```
SELECT DISTINCT institution
FROM learner_data
WHERE institution LIKE '%\%%' OR institution LIKE '%@%' OR institution
```

	institution text
1	%22i Have Completed My Intermediate From Govt. Ghulam Hussain Hidayatullah Higher Secondary School, Hyderabad
2	@Pir Abdul Qadir Shah Jilani Institute Of Medical Science Gambat
3	+2 B M High School Pinjrawan
4	A Mother%27s Touch Preschool And Daycare
5	A%27bomey Calavi, Cotonou, Republic Of Benin
6	ÃCole Supã©Rieure Panafricaine D%27applique Management (Espam Formation University)
7	Abhyasa Women%27s Degree College
8	Aditya Degree College For Women%27s
9	Aditya Women%27s Degree College

```
LIKE '+%';
```

```
--Query to Identify Similar Names
SELECT institution, COUNT(*)
FROM learner_data
GROUP BY institution
HAVING COUNT(*) > 1
ORDER BY COUNT(*) DESC;
```

--

	institution text	count bigint
1	Unknown	52694
2	Saint Louis University	3291
3	Illinois Institute Of Technology	820
4	University Of Lagos	708
5	University Of Ghana	570
6	University Of Ilorin	556
7	Kwame Nkrumah University Of Science And Technology	515
8	University Of Ibadan	492
9	University Of Benin	460
10	Obafemi Awolowo University	387

Query to view all institutions starting with numbers

```
SELECT DISTINCT institution
FROM learner_data
WHERE institution ~ '^[0-9]';
```

	institution text
1	02 Academy
2	1. Hussaini Adamu Federal Polytechnic Kazaure. 2. Collage Of Health Sciences And Technology Tsa...
3	10alytics Academy, Uk
4	10th Ramadan University
5	12th Grade
6	2 B M High School Pinjrawan
7	2023 Tomtom, Openstreetmap Sri Venkateswara College Of Engineering
8	3.051e13
9	30th June Educational Group
10	32ad Academy

```
--Query to view total instition name beginning with 'I '
SELECT major
FROM learner_data
WHERE major ~* '^I\s'
ORDER BY major;
```

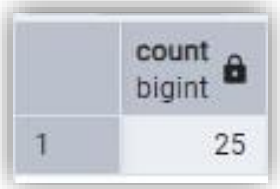
--

	major text
1	I Am A Practicing Lawyer
2	I Am A Senior High School Graduated with Honors From General Academic Stra...
3	I Am Certified Drone Pilot From DGCA
4	I Am Currently Not in School
5	I Am Doing Finding Job
6	I Am Doing Microsoft Power BI Course
7	I Am Major
8	I Am Working As A Education Counsellor Now
9	I Dont Have Any Idea About This
10	I Got An Admission for Doing MS
11	I Have Completed My MBBS
12	I Have No Major
13	I Love Things Concerning Art and Business
14	I Want To Acquire Skills in Data Science

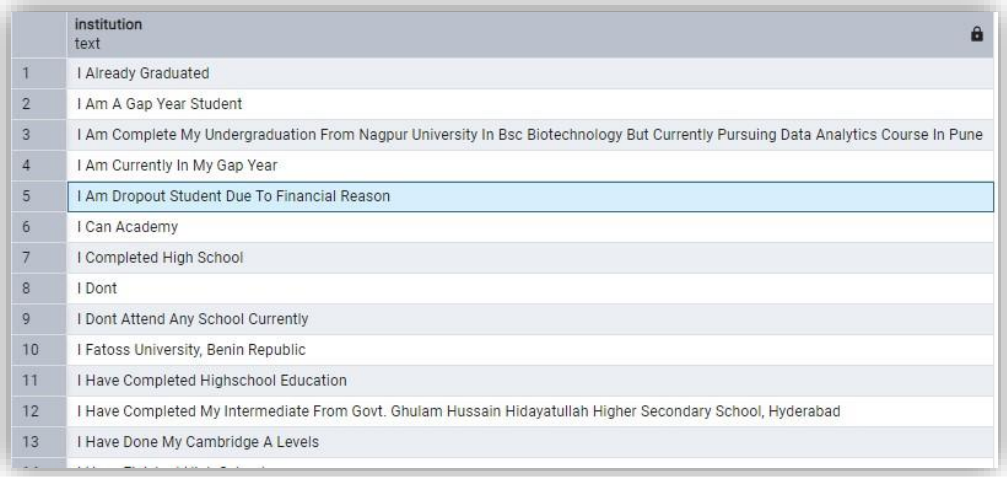
Query to view total institution name beginning with 'I '

```
SELECT COUNT(*)
FROM learner_data
WHERE institution ~* '^I\s';
```

--



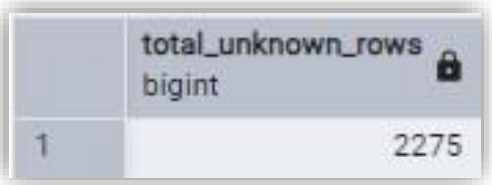
	count bigint
1	25



	institution text
1	I Already Graduated
2	I Am A Gap Year Student
3	I Am Complete My Undergraduation From Nagpur University In Bsc Biotechnology But Currently Pursuing Data Analytics Course In Pune
4	I Am Currently In My Gap Year
5	I Am Dropout Student Due To Financial Reason
6	I Can Academy
7	I Completed High School
8	I Dont
9	I Dont Attend Any School Currently
10	I Fatoss University, Benin Republic
11	I Have Completed Highschool Education
12	I Have Completed My Intermediate From Govt. Ghulam Hussain Hidayatullah Higher Secondary School, Hyderabad
13	I Have Done My Cambridge A Levels

--Query to Count Rows with "Unknown" in Country, Institution, and Major

```
SELECT
    COUNT(*) AS total_unknown_rows
FROM learner_data
WHERE country = 'Unknown'
AND institution = 'Unknown'
AND major = 'Unknown';
```



	total_unknown_rows bigint
1	2275

--START CLEANING NOW.

--Convert to Proper Case (Title Case)

```
UPDATE learner_data
SET institution = INITCAP(institution);
```

--Query to Remove These Values

```
UPDATE learner_data

SET institution = NULL
WHERE institution IN ('-', '----', '..', '...', '.', '(', '(Amu)');
```

--Query to Remove or Nullify Numeric Values

```
UPDATE learner_data

SET institution = NULL

WHERE institution ~ '^[0-9]+$';
```

--Query to Fix Encoded Characters

```
UPDATE learner_data

SET institution = REPLACE(institution, '%22', '"'); -- Replace %22 with '
```

--Query to Remove Special Characters at the Start

```
UPDATE learner_data

SET institution = REGEXP_REPLACE(institution, '^[+@]+', '', 'g')  WHERE

institution ~ '^[+@]+';
```

--Query to NULL those values having email like format(@edu.pk)

```
UPDATE learner_data

SET institution = NULL

WHERE institution LIKE '%\%%' OR institution LIKE '%@%' OR institution LIKE '+%';
```

```
--Query to Convert to Title Case
UPDATE learner_data
SET institution = INITCAP(institution);

--Query to update Merging Variants
UPDATE learner_data
SET institution = 'October University'
WHERE institution IN ('6october University', '6 October University');

--Query to Remove Placeholder Values
UPDATE learner_data
SET institution = NULL
WHERE institution ILIKE ANY (ARRAY['Non', 'Nil', 'Na', 'N/A', 'None', 'Noun', 'No']);

--Query to Remove Special Characters
UPDATE learner_data
SET institution = REGEXP_REPLACE(institution, '[^a-zA-Z0-9\s\(\)\-.,]', '', 'g') WHERE
institution IS NOT NULL;

--Query to Set Invalid Institutions to NULL
UPDATE learner_data
SET institution = NULL
WHERE institution ILIKE ANY (ARRAY[
    '12 Grade', '147thl', '1st Year College', '3.051e+13', 'Ã‰Conomie Et Gestion', 'Ã‘It']);

--Query to Trim Whitespace
UPDATE learner_data
SET institution = TRIM(institution)
WHERE institution IS NOT NULL;
```

--Query to Convert to Title Case

```
UPDATE learner_data

SET institution = INITCAP(institution)
WHERE institution IS NOT NULL;
```

--Query to Set Corrupt Values to NULL

```
UPDATE learner_data

SET institution = NULL

WHERE institution ~ '^[[:space:]]*$'
   OR institution ~ '[^\w\s.,()\-\&]';
```

--Query to Replace NULLs with "Unknown"

```
UPDATE learner_data

SET institution = 'Unknown'
WHERE institution IS NULL;
```

--Query to Convert Invalid Institutions to "Unknown"

```
UPDATE learner_data

SET institution = 'Unknown'
WHERE institution ~ '^[0-9]';
```

--Query to update all institutions that have less than 4 characters.

```
UPDATE learner_data

SET institution = 'Unknown'

WHERE LENGTH(institution) < 4;
```

--Query to view the changes SELECT

```
DISTINCT institution

FROM learner_data

ORDER BY institution;
```

```
--Query to Delete Institutions Starting with "I "
DELETE FROM learner_data
WHERE institution ~* '^I\s';

--Query to Delete Rows having 3 columns as Unknown
DELETE FROM learner_data
WHERE country = 'Unknown'
AND institution = 'Unknown'
AND major = 'Unknown';

--Query to Delete major Starting with "I "
DELETE FROM learner_data
WHERE major ~* '^I\s';

--Query to delete all majors exceeding 100 character length
DELETE FROM learner_data
WHERE LENGTH(major) > 100;

--Query to delete all institution exceeding 100 character length
DELETE FROM learner_data
WHERE LENGTH(institution) > 100;

-- Query to Update country, major, institution, and degree to Lowercase
UPDATE learner_data
SET country = LOWER(country),
major = LOWER(major),    institution
= LOWER(institution),
          degree = LOWER(degree);
```

```
--VERIFICATION:
--Query to Check the Total Row Count

SELECT COUNT(*) FROM learner_data;
```

	count bigint
1	126939

--Query to Check for Any Remaining NULL Values

```
SELECT COUNT(*)
FROM learner_data
WHERE country IS NULL OR major IS NULL OR institution IS NULL;
```

	count bigint
1	0

--Query to Verify Data Standardization

```
SELECT country, major, institution
FROM learner_data
WHERE country LIKE ' %' OR major LIKE ' %' OR institution LIKE ' %'
LIMIT 10;
```

	country character varying (50)	major text	institution text

--Query to Check If All Country, Major, and Institution Names Are Lowercase

```
SELECT * FROM learner_data
WHERE country ~ '[A-Z]' OR major ~ '[A-Z]' OR institution ~ '[A-Z]'
LIMIT 10;
```

	learner_id [PK] uuid	country character varying (50)	degree text	institution text	major text

--Query to Identify Duplicate Records in learner_id

```
SELECT learner_id, COUNT(*)
```

```
FROM learner_data
GROUP BY learner_id
HAVING COUNT(*) > 1;
```

	learner_id [PK] uuid	count bigint

```
--Query to view the final table
SELECT * FROM learner_data limit 10;
```

	learner_id [PK] uuid	country character varying (50)	degree text	institution text	major text
1	aeaee444-9c38-43d8-8c42-6213fdea6e04	pakistan	bachelor	b school	mineral and mining engineering
2	74cd5829-7040-41c0-9940-7fffdf2e16c1	nigeria	graduate	unknown	data science and artificial intelligen...
3	651294e6-924d-41e0-a6a6-1359afd66d58	ghana	other	unknown	data analytics
4	75d9e2f1-d3ad-4861-af4c-a7fa38195518	egypt	graduate	unknown	pharmacy and pharmacology
5	58f4541a-32a1-475b-b64e-3db27fabaa0b	kenya	other	unknown	data science
6	58da6a9e-7ab8-4450-91d4-68e90615e319	egypt	graduate	unknown	management
7	baccd5de-a007-433f-8e87-6ba4f6671fd7	philippines	graduate	unknown	electronics and communication
8	361707c7-a84d-44f1-ba04-5c8930f433b6	india	graduate	unknown	information technology
9	bad5b765-23b5-4e35-823e-27697c27d3e2	south africa	bachelor	unknown	data science
10	68348918-7695-4318-ba4a-a7abdd858a...	vietnam	other	unknown	business and management studies

3. 2. Opportunity Dataset:

```
---- Query to View Null Values in Each Column
SELECT
        SUM(CASE WHEN opportunity_id IS NULL THEN 1 ELSE 0 END) AS
        missing_opportunity_id,
        SUM(CASE WHEN opportunity_name IS NULL THEN 1 ELSE 0 END) AS
        missing_opportunity_name,
          SUM(CASE WHEN category IS NULL THEN 1 ELSE 0 END) AS missing_category,
        SUM(CASE WHEN opportunity_code IS NULL THEN 1 ELSE 0 END) AS
        missing_opportunity_code,
        SUM(CASE WHEN tracking_questions IS NULL THEN 1 ELSE 0 END) AS
        missing_tracking_questions
FROM opportunity_data;
```

	missing_opportunity_id bigint	missing_opportunity_name bigint	missing_category bigint	missing_opportunity_code bigint	missing_tracking_questions bigint
1	0	0	0	0	69

--Query to view opportunity name exceeding 50 character

```
SELECT opportunity_id, opportunity_name, LENGTH(opportunity_name) AS name_length
FROM opportunity_data
WHERE LENGTH(opportunity_name) > 50
ORDER BY name_length DESC;
```

	opportunity_id [PK] character varying (30)	opportunity_name text	name_length integer
1	0000000010X2B85MQE0B6RNEPS	Unlock the Secrets of Chinaâ□□s Ecological Diversity: A Cultural Adventure Await...	82
2	000000000GCKFV5K6Q8FWGFH...	Choosing and Planning for Your Major + Career Exploration Workshops ? In-pers...	79
3	0000000010KMTAJJZCPRAHGW...	The Happiness Project â□□ Exploring the Science and Philosophy of Well-Being	76
4	0000000010T488998MJ34E81AH	Public Speaking Workshop - Crafting and Delivering Persuasive Speeches	70
5	000000000GZKMJYENG510PPM...	Humanizing Technology: An Introduction to User Experience ? In-person	69
6	0000000010ZYTHAM54QTMB69...	Crafting a Powerful Statement of Purpose (SOP) â□□ Telling Your Story	69
7	0000000010NQ273YXVEPW10W...	Mastering Cybersecurity: Safeguarding Confidentiality and Integrity	67
8	0000000010VHKDXBZ03CP36XNR	Health Equity: Addressing Medical Needs in Low-Income Communities	65
9	000000000GN2A0AY7XK8C5FZPP	Career Essentials: Getting Started with Your Professional Journey	65
10	0000000010DN4HDJ9CATZD75...	Navigating Urban Skies: Drone Solutions for Traffic Optimization	64

--Query to view rows having special character

```
SELECT opportunity_name
FROM opportunity_data
WHERE opportunity_name LIKE '%27%';
```

	opportunity_name text
1	Life Beyond Saint Louis University27s Campus
2	Work 2030 Adapting to Tomorrow%27s Workplace
3	The Creator27s Journey: Turning Content into Reven...
4	Pepagora27s Market Mastery Challenge

--START CLEANING

--Query to drop the tracking question column as it is not necessary for analysis

```
ALTER TABLE opportunity_data
DROP COLUMN tracking_questions;
```

```
--Query to remove Leading and Trailing Spaces in All String Columns
UPDATE opportunity_data   SET
   opportunity_name = TRIM(opportunity_name),
category = TRIM(category),
   opportunity_code = TRIM(opportunity_code);
```

```
-- Query to Standardize category Column
UPDATE opportunity_data
SET category = 'Internship'
WHERE category IN ('internship', 'Internships', 'Intern-ship');
```

```
--Query to remove Special Characters from opportunity_name
UPDATE opportunity_data
SET opportunity_name = REGEXP_REPLACE(opportunity_name, '[^a-zA-Z0-9\s\(\)\-.,]',
'')
WHERE opportunity_name IS NOT NULL;
```

```
--Query to delete rows with Non-ASCII character
DELETE FROM opportunity_data
WHERE opportunity_name ~ '[^\x20-\x7E]';
```

```
--Query to delete row with opportunity name more than 70 characters
DELETE FROM opportunity_data
WHERE LENGTH(opportunity_name) > 70;
```

```
--Query to replace unnecessary values
UPDATE opportunity_data
SET opportunity_name = REPLACE(opportunity_name, '%27', '''');
UPDATE opportunity_data
SET opportunity_name = REGEXP_REPLACE(opportunity_name, '27s', 's', 'g');
```

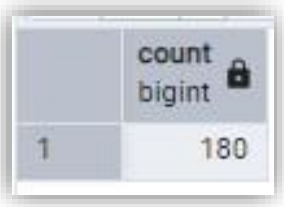
--Query to Remove the ? In-person Text

```
UPDATE opportunity_data
SET opportunity_name = REGEXP_REPLACE(opportunity_name, '\? In-person', '', 'gi');
```

--VERIFICATION:

--Query to check the total rows now

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



	count bigint
1	180

--Query to view rows with Non-ASCII character

```
SELECT COUNT(*)
FROM opportunity_data
WHERE opportunity_name ~ '[^\x20-\x7E]';
```

	count bigint
1	0

--Query to view the final table

SELECT * FROM opportunity_data limit 10;

	opportunity_id [PK] character varying (30)	opportunity_name text	category character varying (50)	opportunity_code character varying (20)
1	0000000010BDV2YMK8R1RN7K...	Accountant	Career	AJENF5Z
2	000000000G127E8VYE08TXBT6X	Choosing and Planning for Your Major	Event	E501873
3	000000000G2PB6VB4ANR28CV...	The Financial Article Writing Competition Te...	Competition	M523594
4	000000000G4AM4J9NBMPK3TJ...	Entrepreneurship and Innovation	Internship	I289641
5	000000000G4F19XBEXPWKS8F...	Statement of Purpose (SOP) Writing Worksh...	Event	E258709
6	000000000G4KVEP36NNJR5YW...	Project Management	Internship	I584159
7	000000000G8BW90E86ARRKM3...	Cybersecurity Defensive Hacking	Internship	I155449
8	000000000G8JG2FEA12SVNXX...	Esports and Game Design	Internship	I860340
9	000000000G95BD07NB0181K0...	Data Visualization	Internship	I660879
10	000000000GBZ5VRTC3YS9T716...	Data Visualization	Internship	I755008

---

3. 3. Cohort Dataset:

-- Query for finding Null Values in Cohort Data

```
SELECT
   SUM(CASE WHEN cohort_code IS NULL THEN 1 ELSE 0 END) AS missing_cohort_code,
   SUM(CASE WHEN start_date IS NULL THEN 1 ELSE 0 END) AS missing_start_date,
   SUM(CASE WHEN end_date IS NULL THEN 1 ELSE 0 END) AS missing_end_date,
   SUM(CASE WHEN size IS NULL THEN 1 ELSE 0 END) AS missing_size
FROM cohort_data;
```

	missing_cohort_code bigint	missing_start_date bigint	missing_end_date bigint	missing_size bigint
1	0	0	0	0

--Query for checking for Duplicates in Cohort Codes (Primary Key)

```
SELECT cohort_code, COUNT(*)
FROM cohort_data
GROUP BY cohort_code
HAVING COUNT(*) > 1;
```

	cohort_code [PK] character varying (20)	count bigint

--Query to identify outlier in cohort sizes

```
SELECT MIN(size) AS min_size,
    MAX(size) AS max_size,
    PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY size) AS q1,
    PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY size) AS median,
    PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY size) AS q3
FROM cohort_data
```

	min_size integer	max_size integer	q1 double precision	median double precision	q3 double precision
1	3	100000	500	800	1500

--START CLEANING

--Query for standardizing Textual Data

```
UPDATE cohort_data
SET cohort_code = LOWER(TRIM(cohort_code));
```

--Query to Replace Outliers with Median Value

```
UPDATE cohort_data
SET size = 800
WHERE size > 3000;
```

--VERIFICATION

--Query for applying data Integrity Check

```
SELECT
   COUNT(DISTINCT cohort_code) AS unique_cohort_codes,
   COUNT(*) AS total_rows
FROM cohort_data;
```

	unique_cohort_codes bigint	total_rows bigint
1	639	639

--Query for checking Data Types

```
SELECT column_name, data_type
```

```
FROM information_schema.columns
WHERE table_name = 'cohort_data';
```

	column_name name	data_type character varying
1	start_date	timestamp without time zone
2	end_date	timestamp without time zone
3	size	integer
4	cohort_code	character varying

```
--Query to verify sizes changes
SELECT DISTINCT size, COUNT(*)
FROM cohort_data
GROUP BY size
ORDER BY size DESC;
```

	size integer	count bigint
1	2500	1
2	1800	9
3	1780	1
4	1771	1
5	1730	1
6	1700	1
7	1600	39
8	1594	1
9	1580	1
10	1550	1

```
--Query to view final table  select
* from cohort_data limit 10;
```

	cohort_code [PK] character varying (20)	start_date timestamp without time zone	end_date timestamp without time zone	size integer
1	bsaem4q	2024-02-05 04:40:00	2024-03-05 11:26:40	800
2	br3l6ku	2023-07-24 08:40:00	2023-08-28 07:33:20	1000
3	bxomkww	2023-05-08 09:26:40	2023-06-09 08:06:40	300
4	b456514	2023-04-03 10:33:20	2023-05-01 21:53:20	1500
5	b328821	2023-01-16 11:20:00	2023-02-20 21:20:00	1000
6	b0vcb0f	2022-09-23 04:40:00	2022-09-23 04:40:00	40
7	b883644	2022-07-01 09:33:20	2022-07-31 09:00:00	1500
8	b759124	2022-10-10 07:46:40	2022-11-15 21:33:20	1500
9	b809432	2022-07-15 18:00:00	2022-07-15 18:00:00	143
10	b495792	2022-07-15 18:00:00	2022-07-15 20:46:40	1771

---

3. 4. Marketing Dataset:

--Query to Select 10 rows

```
SELECT * FROM marketing_data LIMIT 10;
```

--Total Count (141)

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

-- Check for Missing Values in marketing_data

```
SELECT
  SUM(CASE WHEN campaign_name IS NULL THEN 1 ELSE 0 END) AS missing_campaign_name,
  SUM(CASE WHEN reach IS NULL THEN 1 ELSE 0 END) AS missing_reach,
  SUM(CASE WHEN outbound_clicks IS NULL THEN 1 ELSE 0 END) AS missing_outbound_clicks,
  SUM(CASE WHEN landing_page_views IS NULL THEN 1 ELSE 0 END) AS missing_landing_page_views,
  SUM(CASE WHEN result_type IS NULL THEN 1 ELSE 0 END) AS missing_result_type,
  SUM(CASE WHEN results IS NULL THEN 1 ELSE 0 END) AS missing_results,
  SUM(CASE WHEN cost_per_result IS NULL THEN 1 ELSE 0 END) AS missing_cost_per_result,
```

```
   SUM(CASE WHEN amount_spent_aed IS NULL THEN 1 ELSE 0 END) AS
missing_amount_spent_aed,
   SUM(CASE WHEN cpc IS NULL THEN 1 ELSE 0 END) AS missing_cpc
FROM marketing_data;
```

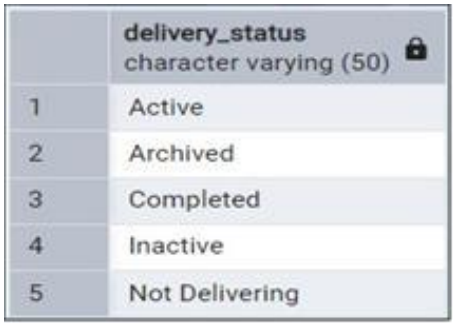
	missing_campaign_name bigint	missing_reach bigint	missing_outbound_clicks bigint	missing_landing_page_views bigint	missing_result_type bigint	missing_results bigint	missing_cost_per_result bigint	missing_amount_spent_aed bigint	missing_cpc bigint
1	0	0	2	0	0	0	0	0	0

```
-- Setting Numeric NULLs to 0
UPDATE marketing_data
SET outbound_clicks = 0
WHERE outbound_clicks IS NULL;
```

```
SELECT
   SUM(CASE WHEN outbound_clicks IS NULL THEN 1 ELSE 0 END) AS
missing_outbound_clicks
FROM marketing_data;
```

	missing_outbound_clicks bigint
1	0

```
-- Standardizing Delivery Status
UPDATE marketing_data SET delivery_status = 'Active' WHERE delivery_status IN
('active');
UPDATE marketing_data SET delivery_status = 'Archived' WHERE delivery_status IN
('archived');
UPDATE marketing_data SET delivery_status = 'Completed' WHERE delivery_status IN
('completed', 'recently_completed');
UPDATE marketing_data SET delivery_status = 'Inactive' WHERE delivery_status IN
('inactive');
UPDATE marketing_data SET delivery_status = 'Not Delivering' WHERE delivery_status IN
('not_delivering');


SELECT DISTINCT delivery_status FROM marketing_data ORDER BY delivery_status;
```



	delivery_status character varying (50)
1	Active
2	Archived
3	Completed
4	Inactive
5	Not Delivering

-- Standardizing Result Type

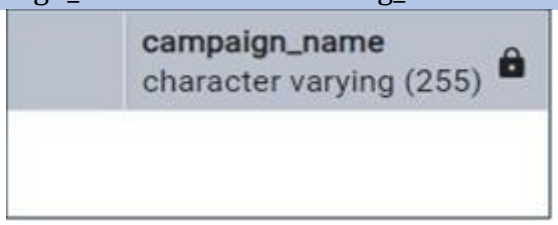
```
UPDATE marketing_data SET result_type = 'Estimated Ad Recall' WHERE result_type = 'Estimated ad recall lift (people)';

UPDATE marketing_data SET result_type = 'Reach' WHERE result_type = 'Reach';

UPDATE marketing_data SET result_type = 'ThruPlay' WHERE result_type = 'ThruPlay';

UPDATE marketing_data SET result_type = 'Website Applications' WHERE result_type = 'Website applications submitted';

UPDATE marketing_data SET result_type = 'Website Leads' WHERE result_type = 'Website leads';


SELECT DISTINCT result_type FROM marketing_data ORDER BY result_type;
```

	result_type character varying (100)
1	Estimated Ad Recall
2	Reach
3	ThruPlay
4	Website Applications
5	Website Leads

-- Cleaning Campaign Names Ending with 'copy%'

```
UPDATE marketing_data

SET campaign_name = TRIM(REGEXP_REPLACE(campaign_name, 'copy[\s\d]*$', '', 'gi'))
```

```
WHERE campaign_name ILIKE '%copy%';

-- Remove Special Characters from Start and End of Campaign Name

UPDATE marketing_data

SET campaign_name = REGEXP_REPLACE(campaign_name, '^[^a-zA-Z0-9]+|[^a-zA-Z0-9]+$', '', 'g')

WHERE campaign_name ~ '^[^a-zA-Z0-9]|[^a-zA-Z0-9]$';

SELECT DISTINCT campaign_name FROM marketing_data ORDER BY campaign_name;
```



```
-- Trimming extra spaces
UPDATE marketing_data
SET campaign_name = TRIM(campaign_name),
ad_account_name = TRIM(ad_account_name),
delivery_status = TRIM(delivery_status),     result_type
= TRIM(result_type);

-- Handling Negative or Unrealistic Values
UPDATE marketing_data
SET reach = GREATEST(reach, 0),
   outbound_clicks = GREATEST(outbound_clicks, 0),
landing_page_views = GREATEST(landing_page_views, 0),     results
= GREATEST(results, 0),
   cost_per_result = GREATEST(cost_per_result, 0),
amount_spent_aed = GREATEST(amount_spent_aed, 0),     cpc
= GREATEST(cpc, 0);

-- Removing duplicates
DELETE FROM marketing_data
WHERE ctid NOT IN (
   SELECT MIN(ctid)
   FROM marketing_data
   GROUP BY ad_account_name, campaign_name, reporting_starts, reporting_ends
);
-- Check if duplicates exist
SELECT campaign_name, COUNT(*) AS duplicate_count
FROM marketing_data
GROUP BY campaign_name
HAVING COUNT(*) > 1
ORDER BY duplicate_count DESC;


SELECT ad_account_name, campaign_name
FROM marketing_data m1
```

```
WHERE EXISTS (
   SELECT 1 FROM marketing_data m2
   WHERE m1.campaign_name = m2.campaign_name
   AND m1.ctid <> m2.ctid
);
```

	ad_account_name character varying (255)	campaign_name character varying (255)
1	SLU	EVENT: Social Impact Initiative
2	RIT	EVENT: Social Impact Initiative
3	Brand Awareness	Dec | Materclasses | Block Chain Essentials Masterclass
4	SLU	Dec | Materclasses | Block Chain Essentials Masterclass

```
-- Standarizing the Text Columns
UPDATE marketing_data
SET campaign_name = INITCAP(campaign_name),  -- Converts to "Title Case"
ad_account_name = UPPER(ad_account_name),  -- Converts to "UPPERCASE"
delivery_status = LOWER(delivery_status), -- Converts to "lowercase"      result_type =
INITCAP(result_type);


-- Final Validation
SELECT DISTINCT * FROM marketing_data LIMIT 10;
```

3. 5. Learner Opportunity Dataset:

```
-- Query to Check total number of records
SELECT COUNT(*) AS total_rows FROM learner_opportunity_data;
```

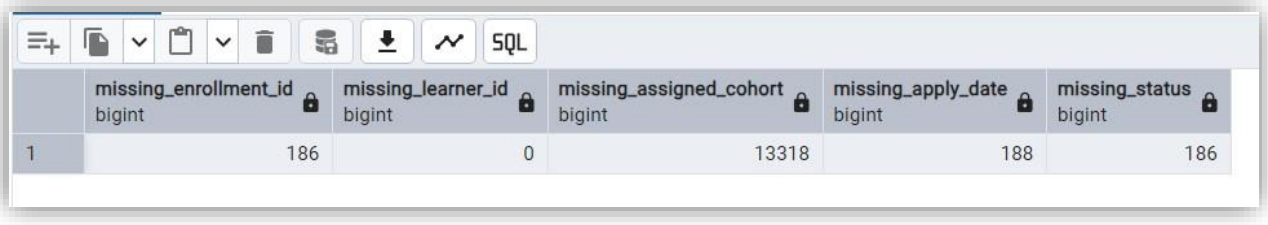
	total_rows bigint
1	113602

```
-- Query to Check for Duplicates (Complete Duplicate Rows)
SELECT enrollment_id, learner_id, assigned_cohort, apply_date, status, COUNT(*)
FROM learner_opportunity_data
GROUP BY enrollment_id, learner_id, assigned_cohort, apply_date, status
```

	enrollment_id uuid	learner_id character varying (30)	assigned_cohort character varying (20)	apply_date timestamp without time zone	status integer	count bigint

```
HAVING COUNT(*) > 1;
```

```
-- Query to Identify Missing (NULL) Values in Each Column
SELECT
    SUM(CASE WHEN enrollment_id IS NULL THEN 1 ELSE 0 END) AS missing_enrollment_id,
    SUM(CASE WHEN learner_id IS NULL THEN 1 ELSE 0 END) AS missing_learner_id,
    SUM(CASE WHEN assigned_cohort IS NULL THEN 1 ELSE 0 END) AS missing_assigned_cohort,
    SUM(CASE WHEN apply_date IS NULL THEN 1 ELSE 0 END) AS missing_apply_date,
    SUM(CASE WHEN status IS NULL THEN 1 ELSE 0 END) AS missing_status
FROM learner_opportunity_data;
```



	missing_enrollment_id bigint	missing_learner_id bigint	missing_assigned_cohort bigint	missing_apply_date bigint	missing_status bigint
1	186	0	13318	188	186

```
-- Check for Incorrect `enrollment_id` Format (UUID Validation)
```

```
SELECT enrollment_id
FROM learner_opportunity_data
WHERE enrollment_id IS NULL;

-- Validate `apply_date` Format and Consistency
-- a) Identify NULL or invalid `apply_date` values
SELECT * FROM learner_opportunity_data
WHERE apply_date IS NULL;

-- b) Find `apply_date` values in the future (potential errors)
UPDATE learner_opportunity_data
SET apply_date = NULL
```


```
WHERE apply_date > NOW();

-- c) Find `apply_date` values that are too old (before 2000)
SELECT * FROM learner_opportunity_data
WHERE apply_date < '2000-01-01';
```


```
-- d) Check for outliers in `apply_date` by analyzing distribution
SELECT DATE_TRUNC('month', apply_date) AS month, COUNT(*)
FROM learner_opportunity_data
GROUP BY month
ORDER BY month;
```

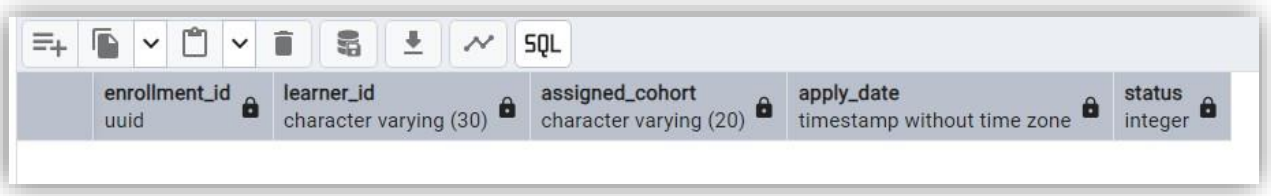
	month timestamp without time zone	count bigint
1	2022-06-01 00:00:00	1
2	2022-07-01 00:00:00	5
3	2022-08-01 00:00:00	511
4	2022-09-01 00:00:00	1046
5	2022-10-01 00:00:00	1000
6	2022-11-01 00:00:00	424
7	2022-12-01 00:00:00	337
8	2023-01-01 00:00:00	598
9	2023-02-01 00:00:00	621
10	2023-03-01 00:00:00	1650
11	2023-04-01 00:00:00	1597
12	2023-05-01 00:00:00	2746
13	2023-06-01 00:00:00	6519
14	2023-07-01 00:00:00	4562
15	2023-08-01 00:00:00	3813
16	2023-09-01 00:00:00	1630
17	2023-10-01 00:00:00	1412
18	2023-11-01 00:00:00	744
19	2023-12-01 00:00:00	1511
20	2024-01-01 00:00:00	2869
21	2024-02-01 00:00:00	2033
22	2024-03-01 00:00:00	8814
23	2024-04-01 00:00:00	6171
24	2024-05-01 00:00:00	7229
25	2024-06-01 00:00:00	1897
26	2024-07-01 00:00:00	1694
27	2024-08-01 00:00:00	4942
28	2024-09-01 00:00:00	6930
29	2024-10-01 00:00:00	6642
30	2024-11-01 00:00:00	7202
31	2024-12-01 00:00:00	6783
32	2025-01-01 00:00:00	10647
33	2025-02-01 00:00:00	8834
34	[null]	188

-- 6. Validate `status` Values (Check for Unexpected Values)

```
SELECT DISTINCT status FROM learner_opportunity_data;
```

	status integer
1	[null]
2	1050
3	1110
4	1070
5	1030
6	1020
7	1080
8	1120
9	1055
10	1040
11	1010

```
-- a) Find invalid status values (e.g., negative values)
UPDATE learner_opportunity_data
SET status = NULL
```




```
WHERE status < 0 OR status IS NULL;
```

```
-- 7. Identify Orphaned `learner_id`s (if it should exist in another table)
SELECT table_name
FROM information_schema.tables
WHERE table_name = 'learners';
```

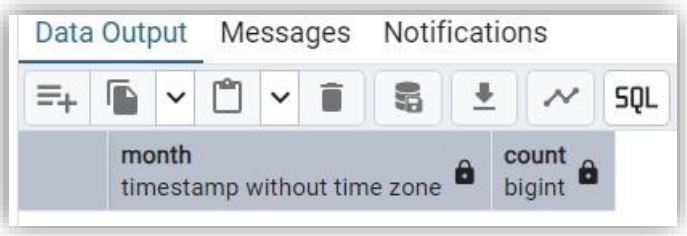
-- 8. Check for Invalid `assigned_cohort` Values

-- a) Find cohorts with unexpected special characters or length > 20

```
UPDATE learner_opportunity_data
SET assigned_cohort = LEFT(REGEXP_REPLACE(assigned_cohort, '[^A-Za-z0-9_-]', '', 'g'), 20)
WHERE assigned_cohort IS NOT NULL;
```

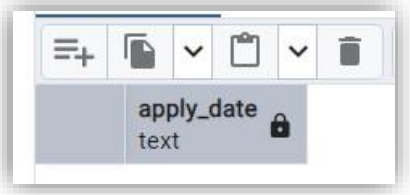
-- 9. Analyze Data Distribution Over Time

```
SELECT DATE_TRUNC('month', apply_date::TIMESTAMP) AS month, COUNT(*)
FROM learner_opportunity_data
GROUP BY month
ORDER BY month;
```



-- 10. Check for Outliers or Anomalies in `apply_date`

```
SELECT apply_date FROM learner_opportunity_data
ORDER BY apply_date DESC
LIMIT 500;
```



Final Query Set for cleaning

--Drop Rows Where enrollment_id is NULL

```
DELETE FROM learner_opportunity_data
WHERE enrollment_id IS NULL;
```

--Handle Missing assigned_cohort Values

```
WITH cohort_mapping AS (
   SELECT learner_id, assigned_cohort,
        COUNT(*) AS cohort_count
   FROM learner_opportunity_data
   WHERE assigned_cohort IS NOT NULL
   GROUP BY learner_id, assigned_cohort
   ORDER BY learner_id, cohort_count DESC
)
UPDATE learner_opportunity_data lo
SET assigned_cohort = cm.assigned_cohort
FROM cohort_mapping cm
WHERE lo.learner_id = cm.learner_id
AND lo.assigned_cohort IS NULL;
```

--setting null values to unknown UPDATE

```
learner_opportunity_data
SET assigned_cohort = 'Unknown'
WHERE assigned_cohort IS NULL;
```

-- Convert apply_date to TIMESTAMP Format

```
ALTER TABLE learner_opportunity_data
ALTER COLUMN apply_date TYPE TIMESTAMP
USING apply_date::TIMESTAMP;
```

--Convert status to INTEGER

```
ALTER TABLE learner_opportunity_data
ALTER COLUMN status TYPE INTEGER USING status::INTEGER;


--deleting missing values in apply_date
DELETE FROM learner_opportunity_data
WHERE apply_date IS NULL;


--the final table
select * from learner_opportunity_data limit 10;
```

	enrollment_id uuid	learner_id character varying (30)	assigned_cohort character varying (20)	apply_date timestamp without time zone	status integer
1	f1f8c1cb-cd2e-40b1-a866-8b3f9359e961	000000000GN2A0AY7XK8C5FZPP	B6MZ4HK	2024-03-30 02:30:15.407	1070
2	08c3ebf1-f27d-4db5-957e-322231fd88...	000000000GN2A0AY7XK8C5FZPP	BSEV9QO	2024-10-07 22:14:24.013	1070
3	ec7cc363-45f0-4f36-819e-004ad86424...	0000000010864WVF3E9R3CXVDW	BN42GTI	2024-10-28 02:22:31.594	1070
4	c8a64974-ffba-4b92-b27b-f735cfe00ffa	0000000010DCSM71JHK7WD6F3R	BUUQU7A	2023-04-23 12:20:26.761	1070
5	47136249-e949-4fa3-bea4-bfe63dbc9c...	0000000010ANNE7X7T1SS2JJPK	BIR2JUR	2025-01-23 18:03:52.363	1070
6	88df3146-b0b4-4f15-a9e0-f6c5a42ec2...	00000000104SZ1BFR638P058YP	BAJRMK1	2024-04-04 16:45:56.852	1055
7	5f03db7c-a8ef-4583-9259-65d190e239...	00000000100JPM3A0WJ85T68M5	BEXFE8O	2024-10-23 16:47:43.562	1070
8	428bb416-0389-41d7-9fbe-aff4798a41...	000000000GSC2DE9P9BZYJM2JN	B167540	2022-10-10 14:25:56.973	1055
9	0f70dfd5-ab92-4fea-9734-d340fb519226	000000000GN2A0AY7XK8C5FZPP	BWAG78I	2025-01-30 02:00:35.535	1070
10	daacee40-916a-43a2-aba9-cc1f725a03...	000000000GNTFT74MZT893VC0G	BHMT70S	2024-08-22 12:28:17.111	1070

3. 6. Cognito Dataset:

```
--Data Distribution and Trends Gender Distribution:
SELECT gender, COUNT(*) AS count
FROM cognito_data
```

```
GROUP BY gender
ORDER BY count DESC;
```

	gender character varying (30)	count bigint
1	Male	49344
2	[null]	42862
3	Female	36515
4	Don%27t want to specify	361
5	Other	96

```
--City Distribution (Top 10):
SELECT city, COUNT(*) AS count
FROM cognito_data
GROUP BY city
ORDER BY count DESC
LIMIT 10;
```

	city character varying (100)	count bigint
1	[null]	42863
2	Lagos	3031
3	Nairobi	2675
4	Accra	2136
5	Karachi	1841
6	Hyderabad	1836
7	Abuja	1532
8	Lahore	1381
9	Ibadan	1221
10	Dhaka	1101

--Identify Missing and Duplicate Values:

```
SELECT
        SUM(CASE WHEN user_id IS NULL THEN 1 ELSE 0 END) AS missing_user_id,
        SUM(CASE WHEN email IS NULL THEN 1 ELSE 0 END) AS missing_email,
         SUM(CASE WHEN gender IS NULL THEN 1 ELSE 0 END) AS missing_gender,
        SUM(CASE WHEN UserCreateDate IS NULL THEN 1 ELSE 0 END) AS
        missing_createdate,
          SUM(CASE WHEN birthdate IS NULL THEN 1 ELSE 0 END) AS missing_birthdate,
        SUM(CASE WHEN city IS NULL THEN 1 ELSE 0 END) AS missing_city,
        SUM(CASE WHEN states IS NULL THEN 1 ELSE 0 END) AS missing_state,
        SUM(CASE WHEN zip IS NULL THEN 1 ELSE 0 END) AS missing_zip
FROM cognito_data;
```

	missing_user_id bigint	missing_email bigint	missing_gender bigint	missing_createdate bigint	missing_birthdate bigint	missing_city bigint	missing_state bigint	missing_zip bigint
1	0	0	42862	0	42862	42863	42864	42867

--Find Duplicate Records (Based on Email)

```
SELECT email, COUNT(*)
FROM cognito_data
GROUP BY email
HAVING COUNT(*) > 1;
```

	email character varying (255)	count bigint
1	mwirigikennedy820@gmail.com	2
2	amaji5295@gmail.com	2
3	nehalmallik00@gmail.com	2
4	amrita02051999@gmail.com	2
5	wediaishaq17@gmail.com	2
6	minahilmumtaz54@gmail.com	2
7	knnabuihe@gmail.com	2
8	daliafouad12345@gmail.com	2
9	bibilola77@gmail.com	2

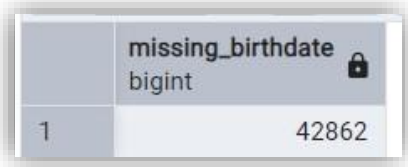
--Handling Email column

```
UPDATE clean_cognito
SET email = NULL
WHERE LENGTH(email) > 50;

UPDATE clean_cognito
SET email = 'Unknown'
WHERE email IS NULL;

--Spot Outliers and Anomalies in city column:
SELECT city
FROM cognito_data
WHERE city IS NOT NULL
ORDER BY LENGTH(city) DESC
LIMIT 10;
```

	city character varying (100)
1	HyderabadHyderabadHyderabadHyderabadHyderabadTemp
2	Sunnyvale Junction by Lokogoma Abuja Nigeria
3	International Islamic University Islamabad
4	Bhod bk taluka dharangaon dist Jalgaon
5	TERNA ENGINEERING COLLEGE NAVI MUMBAI
6	Revathi nagar near karupurayan kovil
7	District Bulandshahr town Aurangabad
8	City of San Jose del monte Bulacan
9	Tehsil Renala khurd District Okara
10	General Santos City South Cotabato

```
--QUERY TO CLEAN GENDER COLOUMN
-- 1 Check how many NULL values exist in gender
SELECT COUNT(*) AS missing_gender
FROM cognito_data
WHERE gender IS NULL;

-- 2 Replace NULLs in gender with 'Not Specified'
```

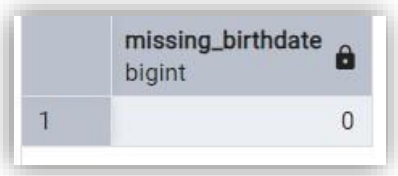
```
UPDATE cognito_data
SET gender = 'Not Specified'
WHERE gender IS NULL;

-- 3 Verify that NULL values are removed from gender
SELECT COUNT(*) AS missing_gender
FROM cognito_data
WHERE gender IS NULL;

-- 4 Prevent future NULL values in gender column
ALTER TABLE cognito_data
ALTER COLUMN gender SET NOT NULL;

--5 Cleaning Don%27t want to specify
UPDATE clean_cognito
SET gender = 'Not Specified'
WHERE gender = 'Don%27t want to specify';

QUERY TO CLEAN BIRTHDATE COLOUMN
-- 1 Check how many NULL values exist in birthdate
SELECT COUNT(*) AS missing_birthdate
FROM cognito_data
WHERE birthdate IS NULL;
```



	missing_birthdate bigint
1	42862

```
-- 2 Replace NULLs in birthdate with '2000-01-01' (Default Birthdate)
UPDATE cognito_data
SET birthdate = '2000-01-01'
WHERE birthdate IS NULL;

-- 3 Verify that NULL values are removed from birthdate
```

```
SELECT COUNT(*) AS missing_birthdate
FROM cognito_data
WHERE birthdate IS NULL;
```



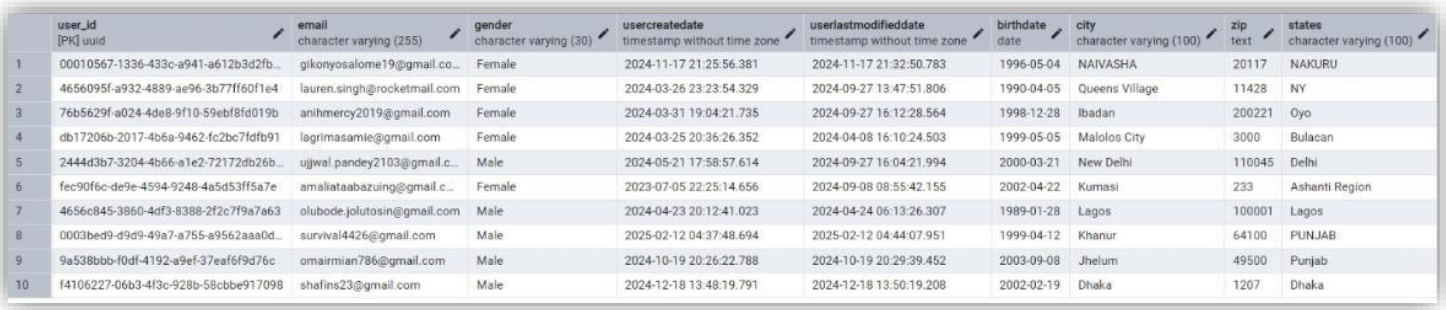
	missing_birthdate bigint
1	0

```
-- 4 Prevent future NULL values in birthdate column
ALTER TABLE cognito_data
ALTER COLUMN birthdate SET NOT NULL;
```

```
--Handling Zip code column
UPDATE clean_cognito
SET zip = NULL
WHERE zip = '00000';

UPDATE clean_cognito
SET zip = TRIM(zip);

UPDATE clean_cognito
SET zip = REGEXP_REPLACE(zip, '\s+', '', 'g')
WHERE zip ~ '[0-9]';

UPDATE clean_cognito
SET zip = REGEXP_REPLACE(zip, '[^0-9\-]', '', 'g')
WHERE zip ~ '[0-9]';

UPDATE clean_cognito
SET zip = NULL
WHERE LENGTH(zip) < 4;

UPDATE clean_cognito
```

```
SET zip = 'Unknown'
WHERE zip IS NULL;

UPDATE clean_cognito
SET zip = NULL
WHERE zip LIKE '-%' OR zip LIKE '%-';

UPDATE clean_cognito
SET zip = 'Unknown'
WHERE zip IS NULL;
```

```
--Viewing the final table select *
from cognito_data limit 10;
```



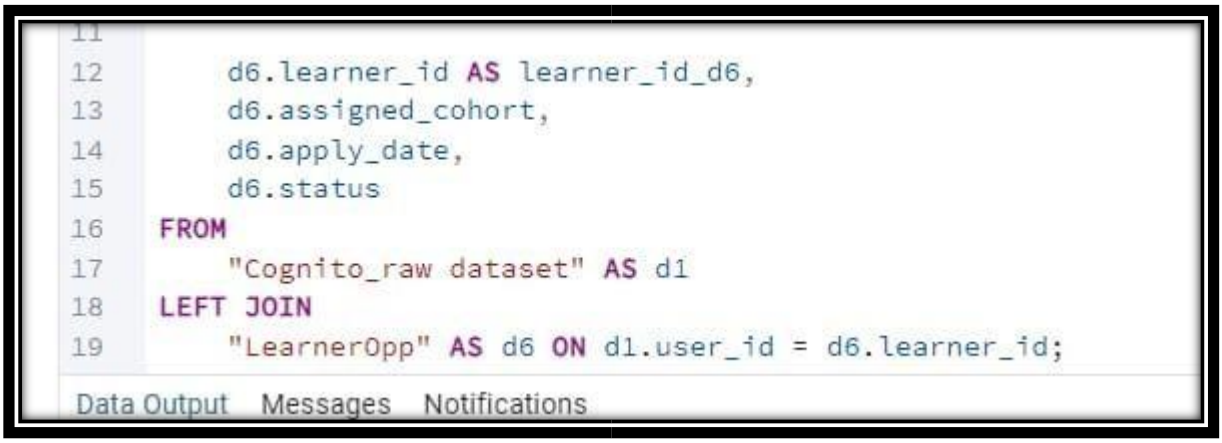
	user_id [PK] uuid	email character varying (255)	gender character varying (30)	usercreatedate timestamp without time zone	userlastmodifieddate timestamp without time zone	birthdate date	city character varying (100)	zip text	states character varying (100)
1	00010567-1336-433c-a941-a612b3d2fb...	gikonyosalome19@gmail.co...	Female	2024-11-17 21:25:56.381	2024-11-17 21:32:50.783	1996-05-04	NAIVASHA	20117	NAKURU
2	4656095f-a932-4889-ae96-3b77ff60f1e4	lauren.singh@rocketmail.com	Female	2024-03-26 23:23:54.329	2024-09-27 13:47:51.806	1990-04-05	Queens Village	11428	NY
3	76b5629f-a024-4de8-9f10-59ebf8fd019b	anihmercy2019@gmail.com	Female	2024-03-31 19:04:21.735	2024-09-27 16:12:28.564	1998-12-28	Ibadan	200221	Oyo
4	db17206b-2017-4b6a-9462-fc2bc7fdfb91	lagrimasamie@gmail.com	Female	2024-03-25 20:36:26.352	2024-04-08 16:10:24.503	1999-05-05	Malolos City	3000	Bulacan
5	2444d3b7-3204-4b66-a1e2-72172db26b...	ujjwal.pandey2103@gmail.c...	Male	2024-05-21 17:58:57.614	2024-09-27 16:04:21.994	2000-03-21	New Delhi	110045	Delhi
6	fec90f6c-de9e-4594-9248-4a5d53ff5a7e	amaliataabazuing@gmail.c...	Female	2023-07-05 22:25:14.656	2024-09-08 08:55:42.155	2002-04-22	Kumasi	233	Ashanti Region
7	4656c845-3860-4df3-8388-2f2c7f9a7a63	olubode.jolutosin@gmail.com	Male	2024-04-23 20:12:41.023	2024-04-24 06:13:26.307	1989-01-28	Lagos	100001	Lagos
8	0003bed9-d9d9-49a7-a755-a9562aaa0d...	survival4426@gmail.com	Male	2025-02-12 04:37:48.694	2025-02-12 04:44:07.951	1999-04-12	Khanur	64100	PUNJAB
9	9a538bbb-f0df-4192-a9ef-37eaf6f9d76c	omairmian786@gmail.com	Male	2024-10-19 20:26:22.788	2024-10-19 20:29:39.452	2003-09-08	Jhelum	49500	Punjab
10	f4106227-06b3-4f3c-928b-58cbbe917098	shafins23@gmail.com	Male	2024-12-18 13:48:19.791	2024-12-18 13:50:19.208	2002-02-19	Dhaka	1207	Dhaka

Chapter 4: Master Table Creation

```
CREATE TABLE MasterTable (
        -- Learner Opportunity Data (Base Table)
            enrollment_id UUID PRIMARY KEY,  -- Unique enrollment identifier (PK)
 learner_id    VARCHAR(30),                    opportunity_id    VARCHAR(30),
assigned_cohort  VARCHAR(20),          apply_date  TIMESTAMP,          status
INTEGER,


        -- Learner Data
        country VARCHAR(50),
        degree TEXT,
        institution TEXT,
        major TEXT,

        -- Opportunity Data
        opportunity_name TEXT,
        category VARCHAR(50),

        -- Cohort Data
        start_date TIMESTAMP,
        end_date TIMESTAMP,

        -- Cognito Data (Demographics)
        user_id UUID,  -- FK linking to clean_cognito (user_id is UUID)
        email VARCHAR(100),       gender VARCHAR(50),
        birthdate DATE,




        city VARCHAR(50),

      -- Foreign Keys (Ensuring proper relationships)
FOREIGN KEY (user_id) REFERENCES clean_cognito(user_id),   FOREIGN
```

```
        KEY (opportunity_id) REFERENCES  clean_opportunity(opportunity_id),
);      FOREIGN KEY (assigned_cohort) REFERENCES clean_cohort(cohort_code)
```

```
SELECT
    d1.user_id,
    d1.email,
    d1.gender,
    d1.user_create_date,
    d1.user_last_modified,
    d1.birthdate,
    d1.city,
    d1.zip,
    d1.state,

    d6.learner_id AS learner_id_d6,
```



```
    d6.learner_id AS learner_id_d6,
    d6.assigned_cohort,
    d6.apply_date,
    d6.status
FROM
    "Cognito_raw dataset" AS d1
LEFT JOIN
    "LearnerOpp" AS d6 ON d1.user_id = d6.learner_id;
```

Data Output Messages Notifications

Showing rows: 1 to 1000 Page No: 1 of 142

	user_id text	email text	gender text	user_create_date time with time zone	user_last_modified time with time zone	birthdate date	city text	zip text
1	4e6f78a9-f9b2-4352-ad22-d43dc46f5ff7	shammasmangat@gmail.com	Male	05:54:05.152000+05:30	11:05:59.728000+05:30	2002-05-31	Malappuram	673314
2	4e9f5cb5-0576-4dbc-b7f5-1fae5f29b2df	samueltasare200@gmail.com	Male	06:38:17.583000+05:30	12:14:36.814000+05:30	1991-03-12	Aboso Prestea HV	233
3	4ea61aa9-17da-4b60-9872-359b8e1e16...	magagbley.ma@gmail.com	Female	21:44:24.851000+05:30	16:08:33.212000+05:30	1991-03-28	Accra	233
4	4eb218c7-467a-470a-9e3e-a2b7bc649e18	najamalialiali449@gmail.com	Male	15:35:35.542000+05:30	18:32:33.778000+05:30	2003-04-04	Islamabad hostel City	45600
5	4ec728db-7d09-4a8e-b1ab-dab3011a55...	franchezcaconcepcion@gmail.com	Female	06:54:37.847000+05:30	06:58:06.099000+05:30	2001-02-06	Makati	1203
6	4f19b0d8-d5f4-463b-b78c-ea2b74db540d	shammah110@gmail.com	Female	23:09:47.442000+05:30	13:49:57.744000+05:30	1999-06-30	Port Harcourt	50002

Chapter 5: ETL Process (Extract, Transform, Load)

```
--Keep Only One Entry Per Enrollment
DELETE FROM clean_learner_opportunity
WHERE enrollment_id IN (
        SELECT enrollment_id
        FROM (
        SELECT enrollment_id,
        apply_date,
                ROW_NUMBER() OVER (PARTITION BY enrollment_id ORDER BY
        apply_date DESC) AS row_num
        FROM clean_learner_opportunity
        ) subquery
        WHERE row_num > 1
);

--Replace Missing assigned_cohort with Unknown
UPDATE clean_learner_opportunity
SET assigned_cohort = 'Unknown'
WHERE assigned_cohort IN (
   SELECT lo.assigned_cohort
   FROM clean_learner_opportunity lo
   LEFT JOIN clean_cohort c ON lo.assigned_cohort = c.cohort_code
WHERE c.cohort_code IS NULL
);

INSERT INTO clean_cohort (cohort_code, start_date, end_date, size)
VALUES ('Unknown', NULL, NULL, 0)
ON CONFLICT (cohort_code) DO NOTHING;
```

```
--inserting learner opportunity data
INSERT INTO MasterTable (enrollment_id, learner_id, assigned_cohort, apply_date,
status) SELECT
   lo.enrollment_id, lo.learner_id, lo.assigned_cohort, lo.apply_date, lo.status FROM
clean_learner_opportunity lo;


--Insert Learner Data
UPDATE MasterTable mt SET
   country = l.country,
degree = l.degree,    institution
= l.institution,    major =
l.major
FROM clean_learner l
WHERE mt.enrollment_id = l.learner_id;


--Insert Opportunity Data
UPDATE MasterTable mt SET
   opportunity_id = lo.learner_id,
   opportunity_name = o.opportunity_name,
category = o.category
FROM clean_learner_opportunity lo
JOIN clean_opportunity o ON lo.learner_id = o.opportunity_id WHERE mt.enrollment_id
= lo.enrollment_id;
```

```
--Insert Cohort Data
UPDATE MasterTable mt SET
   start_date = c.start_date,
end_date = c.end_date
FROM clean_cohort c
WHERE mt.assigned_cohort = c.cohort_code;

--Insert Cognito (User) Data
UPDATE MasterTable mt SET
   user_id = c.user_id,
email = c.email,    gender
= c.gender,    birthdate =
c.birthdate,    city = c.city
FROM clean_cognito c
JOIN clean_learner l ON c.user_id = l.learner_id  WHERE
mt.enrollment_id = c.user_id;

--Dropping Unnecessary Columns and columns having same content
ALTER TABLE MasterTable
DROP COLUMN opportunity_id,
DROP COLUMN assigned_cohort,
DROP COLUMN start_date,
DROP COLUMN end_date,
DROP COLUMN user_id;
```

Chapter 6: Final Validation

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

	count bigint
1	37392

--VIEWING FINAL MASTER TABLE

SELECT * FROM MasterTable;

	enrollment_id [PK] uuid	learner_id character varying (30)	apply_date timestamp without time zon	status integer	country character v	degree text	institution text	major text	opportunity_name text
1	be691596-de19-4d78-9b66-a8e6229698a	0000000000GWQAXC5X45C2MHJ...	2023-06-16 08:11:36.402	1070	india	bachelor	manipal	information technology	Data Visualization E
2	5b61b2e4-d9d0-4094-8f5d-b71ea93ff6ef	0000000010GG17ZMAP4TWQVB...	2025-01-21 04:19:59.176	1070	india	bachelor	muthayammal engineering	artificial intelligence and data science	Data Visualization A
3	7ec1e06f-f112-46dd-8d54-995163877a59	0000000010GG17ZMAP4TWQVB...	2024-08-24 21:16:36.443	1070	nigeria	graduate	federal polytechnic nekede owerri	electrical and electronic engineering	Data Visualization A
4	772f9eda-4cf4-4161-a1fe-8fe680aa4225	0000000010GG17ZMAP4TWQVB...	2025-01-27 13:22:15.247	1070	india	bachelor	amity university, lucknow	data analytics	Data Visualization A
5	582c41cb-4b8e-417b-b554-6a1f47cc6ada	0000000010GG17ZMAP4TWQVB...	2025-01-24 11:53:04.164	1070	india	bachelor	amity university noida	btech	Data Visualization A
6	021a9091-ebe7-4ce3-9cea-b7f5a87c1a69	0000000010VCWKGF64S12KJ9RC	2025-02-24 16:44:20.617	1070	india	graduate	p.p.n collage kanpur	marketing	Dust Extraction Cha
7	77527548-2ac6-4002-80fa-cdc5d428fb0a	00000000104SZ1BFR638P058YP	2024-03-31 11:02:05.444	1055	india	graduate	university of delhi	accounting and finance	Business Developm
8	3466b704-1a43-48b2-b39f-c08555aca4d7	0000000010RQJXA9NNKDEW5RCF	2024-04-05 12:37:14.566	1070	india	bachelor	bannari amman	computer engineering	UX Redesign Challe
9	85f16e55-a9b4-4d66-b121-6c183c2013ae	0000000010SAZXDAE05AN2CGAN	2025-02-20 12:46:36.583	1070	pakistan	graduate	lahore for women university	finance	Project Managemer
10	0bbdab58-5029-4bd4-bf4b-9a598c637f16	0000000000GWQAXC5X45C2MHJ...	2023-07-10 15:46:18.818	1070	nigeria	graduate	nnamdi azikiwe university	politics	Data Visualization E
11	6f76d5c9-3f55-41c8-8448-3c054c92eb35	0000000010GG17ZMAP4TWQVB...	2023-08-15 15:26:27.08	1110	pakistan	bachelor	bahria university	computer science	Data Visualization A
12	b4d15f71-483f-419b-bd7d-a06ea08a2fb7	0000000000GNTFT74MZT893VC0G	2025-01-19 23:40:34.468	1070	kenya	bachelor	daystar university	communication	Digital Marketing Ea
13	d9300915-ba50-4da9-9e2d-f33a2a6c6ffc	0000000010SAZXDAE05AN2CGAN	2025-01-29 14:45:34.075	1070	south a...	bachelor	princess high school	othe	Project Managemer
14	02eadc90-2efc-4372-aadf-1976ddddba40	0000000000GN2A0AY7XK8C5FZPP	2022-11-10 06:54:18.6	1080	nepal	bachelor	islington	computer science and information s...	Career Essentials G
15	feae3822-4e0f-438f-a8b8-7582e778cb11	0000000010SAZXDAE05AN2CGAN	2025-02-13 04:48:44.171	1070	nepal	graduate	tribhuwan university	development studies	Project Managemer
16	da0bfaed-522d-40e8-8e7f-1e962d405949	0000000000GWQAXC5X45C2MHJ...	2024-10-19 16:48:53.576	1070	india	graduate	future group of management	student	Data Visualization E

opportunity_name text	category character varying (50)	email character varying (100)	gender character varying (50)	birthdate date	city character varying (50)
Data Visualization Early Internship	Internship	tpakhare40@gmail.com	Male	2002-09-11	Hyderabad
Data Visualization Associate Early Internship	Internship	gurumaheswarreddy888@gmail.com	Male	2004-04-14	Kadapa
Data Visualization Associate Early Internship	Internship	showeah@gmail.com	Male	1990-03-03	Lagos
Data Visualization Associate Early Internship	Internship	adii.kumar081@gmail.com	Male	2006-08-08	Lucknow
Data Visualization Associate Early Internship	Internship	debshatachoudhury@gmail.com	Male	2005-03-18	New Delhi
Dust Extraction Challenge - Phase 1	Competition	ss7880306696@gmail.com	Female	2001-05-22	Kanpur
Business Development Virtual Internship	Internship	vardanchawla3108@gmail.com	Male	2002-08-31	Rewari
UX Redesign Challenge	Competition	tarunkumar.cs20@bitsathy.ac.in	Male	2002-02-25	Namakkal
Project Management Associate Early Internship	Internship	mahamlatif5@gmail.com	Female	2001-04-04	Lahore
Data Visualization Early Internship	Internship	mrogugua@gmail.com	Male	1992-08-11	Surulere
Data Visualization Associate Early Internship	Internship	ayesha.zb7@gmail.com	Female	2002-10-20	Karachi
Digital Marketing Early Internship	Internship	tarteeanna@gmail.com	Female	1999-06-23	Eldoret
Project Management Associate Early Internship	Internship	kyrospoek@gmail.com	Male	1999-03-20	Johannesburg
Career Essentials Getting Started with Your Professional Journey	Course	uniquemagar07@gmail.com	Male	1998-08-22	Kathmandu
Project Management Associate Early Internship	Internship	thereasong30@gmail.com	Male	1996-02-10	Jhapa
Data Visualization Early Internship	Internship	pandeylokesh946@gmail.com	Male	2007-09-11	Bareilly

Chapter 7: Conclusion

Why This Master Table is Great for Dashboarding?

1. **Key Identifiers**
    - enrollment_id → Primary key ensures each record is unique.
    - learner_id & user_id → Allows linking of learners across datasets.

2. **Demographics & Learner Info**
    - country, degree, institution, major, gender, birthdate, city.

- Visualizations:
  - Bar Charts: Number of learners per country, degree, institution, gender
  - Pie Charts: Distribution of majors ⭘ Age Analysis using birthdate

3. **Enrollment & Status Tracking**
   - apply_date → Can be used to analyze application trends over time
   - status → Allows us to track accepted/rejected learners
   - Visualizations:
     - Time Series Line Graph: Enrollment trends over time ⭘ Bar Chart: Status distribution

4. **Opportunity Analysis**
   - opportunity_name, category
   - Visualizations:
     - Treemap: Number of learners in different opportunity categories ⭘ Top 10 Opportunities with the highest enrollments

5. **Email & Marketing Potential**
   - email → Can be used for communication tracking
   - Possible Metric: Count unique learners with valid email

Why Certain Columns Were Excluded from the Master Table:

In the process of designing the Master Table, we carefully analyzed which columns would be relevant for effective visualization and dashboard creation. Several columns, particularly from the **marketing dataset**, were intentionally excluded to maintain a **clean, meaningful, and efficient dataset**.

1. Exclusion of Marketing Data

We initially considered integrating marketing metrics such as **reach, outbound clicks, CPC, and amount spent**. However, these were omitted for the following reasons:

- **No Direct Linkage to Enrollment Data:** The marketing dataset primarily tracks ad campaign performance, but it lacks a direct connection to specific learners or enrollments.
- **Limited Actionable Insights:** While marketing metrics are valuable for advertisers, they do not contribute significantly to our goal of tracking learner journeys, enrollments, and opportunities.

- **Potential Data Redundancy:** Keeping marketing data would add unnecessary complexity without strong correlations to the enrollment process.

2. Assigned Cohort Column – High Rate of Missing Values

The assigned_cohort column was found to contain a large number of "Unknown" values. This inconsistency made it unreliable for segmentation. Instead of including incomplete and potentially misleading data, we chose to exclude it from major analyses.

3. Dropping Unnecessary Identifiers

Certain IDs and tracking fields (such as cohort_code and some UUIDs) were only used for internal linking but did not add value to data visualization. We focused on keeping only the identifiers that were directly needed for meaningful analysis.

4. Keeping the Master Table Simple & Actionable

Our goal was to create a concise, structured dataset that enables clear, insightful visualizations without cluttering the dashboard. Unnecessary columns would have:

- Increased data processing complexity.
- Made visualizations less intuitive and actionable.
- Introduced inconsistencies due to missing or unstructured data.