Data Quality Exploratory Data Analysis (EDA) Report by Team 2

Dataset Overview

This report reviews key data quality and structural issues identified across datasets related to learners, cohorts, marketing campaigns, and user demographics. The objective is to provide clear observations, risks, and recommended mitigation strategies to improve data integrity and analytic reliability.

Datasets Covered:

- Learner Dataset
- Content and Tracking Data
- Cohort Data
- Marketing Campaign Data
- Data Import Logs
- User Demographic Data

Summary Stats

- The status field in the Learner dataset contains undocumented numeric codes.
- Duplicate learner_id and apply_date combinations detected.
- Multiple cohort assignments found per learner.
- Tracking data shows many NULL entries for questions and inconsistent naming.
- Cohort sizes range from unrealistically large to zero-day durations.
- Marketing campaigns have duplicate names and inactive campaigns with spend.
- Data import suffers from malformed rows and inconsistent ID formats.
- User demographics include missing birthdates, inconsistent gender values, and invalid emails.

Missing & Duplicates

- Duplicate learner_id/apply_date pairs could bias enrollment counts.
- Missing critical fields such as tracking questions and demographic attributes.
- NULL and literal "NULL" values present in NOT NULL fields.
- Duplicate campaign names (e.g., "Copy 3", "Copy 4") affect marketing analysis.
- Malformed rows cause import failures and incomplete records.

Outliers

- Epoch timestamps in UNIX format require conversion.
- Cohort sizes up to 100,000 and durations spanning years flagged.
- Marketing cost per result varies significantly; some campaigns show low engagement despite high reach.
- User demographic zip codes show inconsistent formats.
- Outlier cohort durations and size discrepancies need validation.

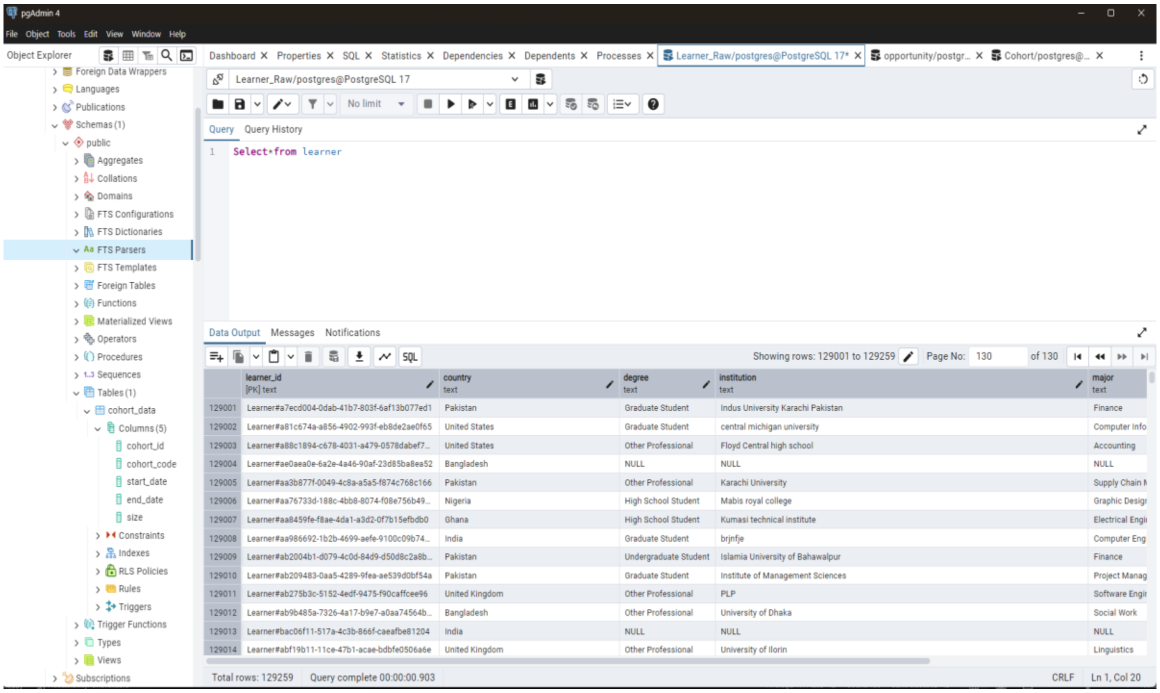
Trend Visuals

- Learner enrollment over time, highlighting duplicates.
- Cohort size and duration distribution.
- Marketing campaign spend vs results over periods.
- Demographic data completeness trends.
- Status code frequency and anomaly detection.

Key Insights

- Enforce data validation rules and create data dictionaries for undocumented codes.
- Implement deduplication and uniqueness constraints on learner and cohort data.
- Standardize naming conventions and clean tracking data entries.
- Set thresholds and validation rules for cohort sizes and durations.
- Optimize marketing campaign lifecycle management to avoid wasted spend.
- Improve import procedures with schema validation and pre-processing.
- Normalize and enrich demographic data to improve analytic accuracy.
- Adopt consistent timestamp and ID formatting standards across datasets.

1. Image of Learner_Raw_Data successfully imported in PostgreSQL



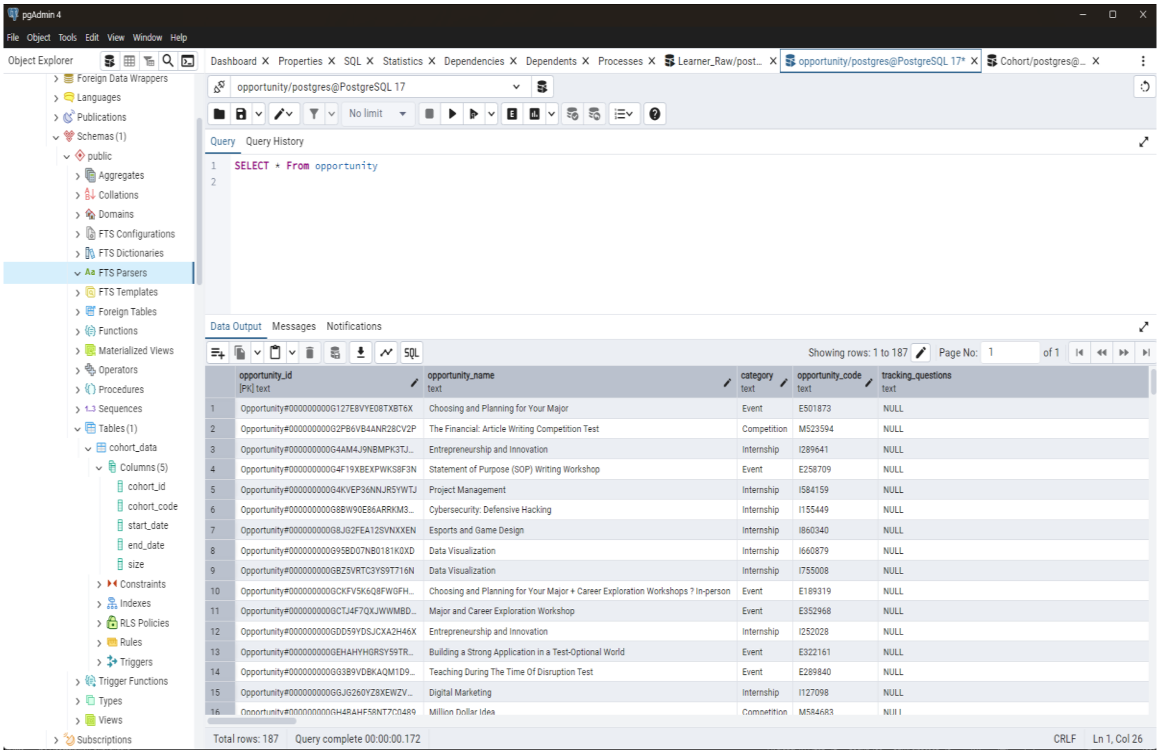
```
Select*from learner
```

Showing rows: 129001 to 129259 Page No: 130 of 130

	learner_id [PK] text	country text	degree text	institution text	major text
129001	Learner#a7ecd004-0dab-41b7-803f-6af13b077ed1	Pakistan	Graduate Student	Indus University Karachi Pakistan	Finance
129002	Learner#a81c674a-a856-4902-993f-eb8de2ae0f65	United States	Graduate Student	central michigan university	Computer Info
129003	Learner#a88c1894-c678-4031-a479-0578dabef7...	United States	Other Professional	Floyd Central high school	Accounting
129004	Learner#ae0aea0e-6a2e-4a46-90af-23d85ba8ea52	Bangladesh	NULL	NULL	NULL
129005	Learner#aa3b877f-0049-4c8a-a5a5-f874c768c166	Pakistan	Other Professional	Karachi University	Supply Chain M
129006	Learner#aa76733d-188c-4bb8-8074-f08e756b49...	Nigeria	High School Student	Mabis royal college	Graphic Desigr
129007	Learner#aa8459fe-f8ae-4da1-a3d2-0f7b15efbdb0	Ghana	High School Student	Kumasi technical institute	Electrical Engi
129008	Learner#aa986692-1b2b-4699-aefe-9100c09b74...	India	Graduate Student	brjnfje	Computer Eng
129009	Learner#ab2004b1-d079-4c0d-84d9-d50d8c2a8b...	Pakistan	Undergraduate Student	Islamia University of Bahawalpur	Finance
129010	Learner#ab209483-0aa5-4289-9fea-ae539d0bf54a	Pakistan	Graduate Student	Institute of Management Sciences	Project Manag
129011	Learner#ab275b3c-5152-4edf-9475-f90caffcee96	United Kingdom	Other Professional	PLP	Software Engir
129012	Learner#ab9b485a-7326-4a17-b9e7-a0aa74564b...	Bangladesh	Other Professional	University of Dhaka	Social Work
129013	Learner#bac06f11-517a-4c3b-866f-caeafbe81204	India	NULL	NULL	NULL
129014	Learner#abf19b11-11ce-47b1-acae-bdbfe0506a6e	United Kingdom	Other Professional	University of Ilorin	Linguistics

Total rows: 129259 Query complete 00:00:00.903 CRLF Ln 1, Col 20

2. Image of Opportunity_Data successfully Imported.



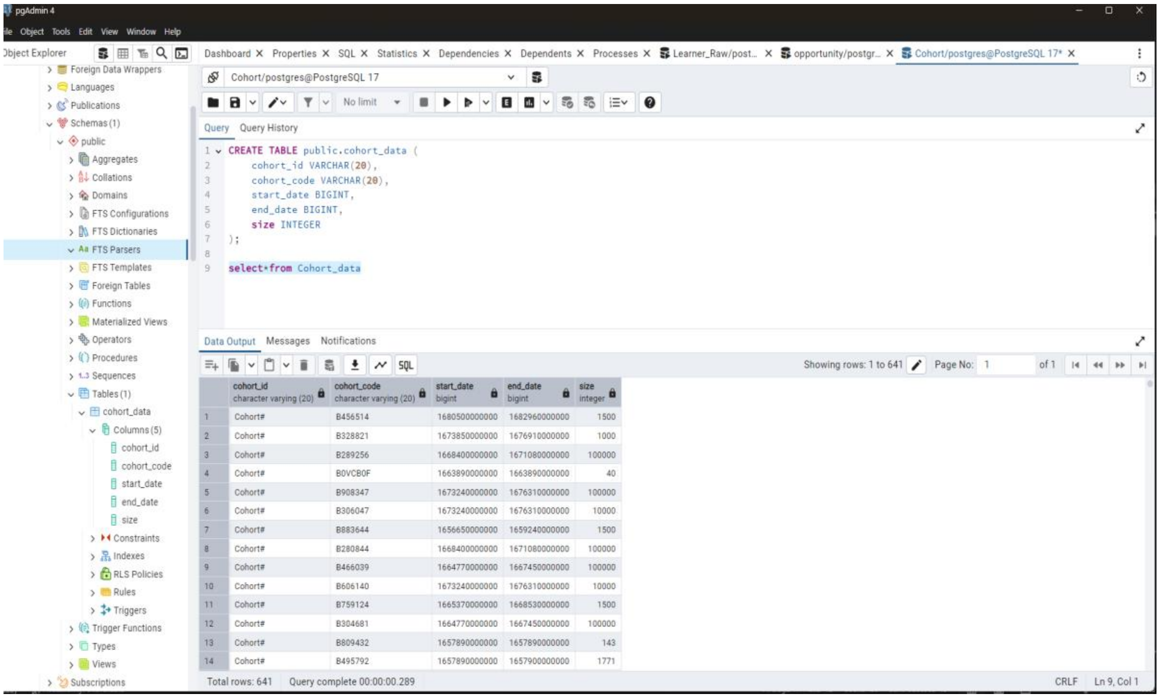
```
SELECT * From opportunity
```

Showing rows: 1 to 187 Page No: 1 of 1

	opportunity_id [PK] text	opportunity_name text	category text	opportunity_code text	tracking_questions text
1	Opportunity#0000000000G127E8VYE08TXBT6X	Choosing and Planning for Your Major	Event	E501873	NULL
2	Opportunity#0000000000G2PB6VB4ANR28CV2P	The Financial: Article Writing Competition Test	Competition	M523594	NULL
3	Opportunity#0000000000G4AM4J9NBMPK3TJ...	Entrepreneurship and Innovation	Internship	I289641	NULL
4	Opportunity#0000000000G4F19XBEXPWKS8F3N	Statement of Purpose (SOP) Writing Workshop	Event	E258709	NULL
5	Opportunity#0000000000G4KVEP36NNJR5YWTJ	Project Management	Internship	I584159	NULL
6	Opportunity#0000000000G8BW90E86ARRKM3...	Cybersecurity: Defensive Hacking	Internship	I155449	NULL
7	Opportunity#0000000000G8JG2FEA12SVNXXEN	Esports and Game Design	Internship	I860340	NULL
8	Opportunity#0000000000G95BD07NB0181K0XD	Data Visualization	Internship	I660879	NULL
9	Opportunity#0000000000GBZ5VRTC3YS9T716N	Data Visualization	Internship	I755008	NULL
10	Opportunity#0000000000GCKFV5K6Q8FWGFH...	Choosing and Planning for Your Major + Career Exploration Workshops ? In-person	Event	E189319	NULL
11	Opportunity#0000000000GCTJ4F7QXJWWMBD...	Major and Career Exploration Workshop	Event	E352968	NULL
12	Opportunity#0000000000GDD59YDSJCXA2H46X	Entrepreneurship and Innovation	Internship	I252028	NULL
13	Opportunity#0000000000GEHAHYHGRSY59TR...	Building a Strong Application in a Test-Optional World	Event	E322161	NULL
14	Opportunity#0000000000GG3B9VDBKAQM1D9...	Teaching During The Time Of Disruption Test	Event	E289840	NULL
15	Opportunity#0000000000GGJG260YZ8XEWZV...	Digital Marketing	Internship	I127098	NULL
16	Opportunity#0000000000GH4BAHF58NT7C0489	Million Dollar Idea	Competition	M584683	NULL

Total rows: 187 Query complete 00:00:00.172 CRLF Ln 1, Col 26

3. Image of Cohort_Data successfully imported



Cohort/postgres@PostgreSQL 17

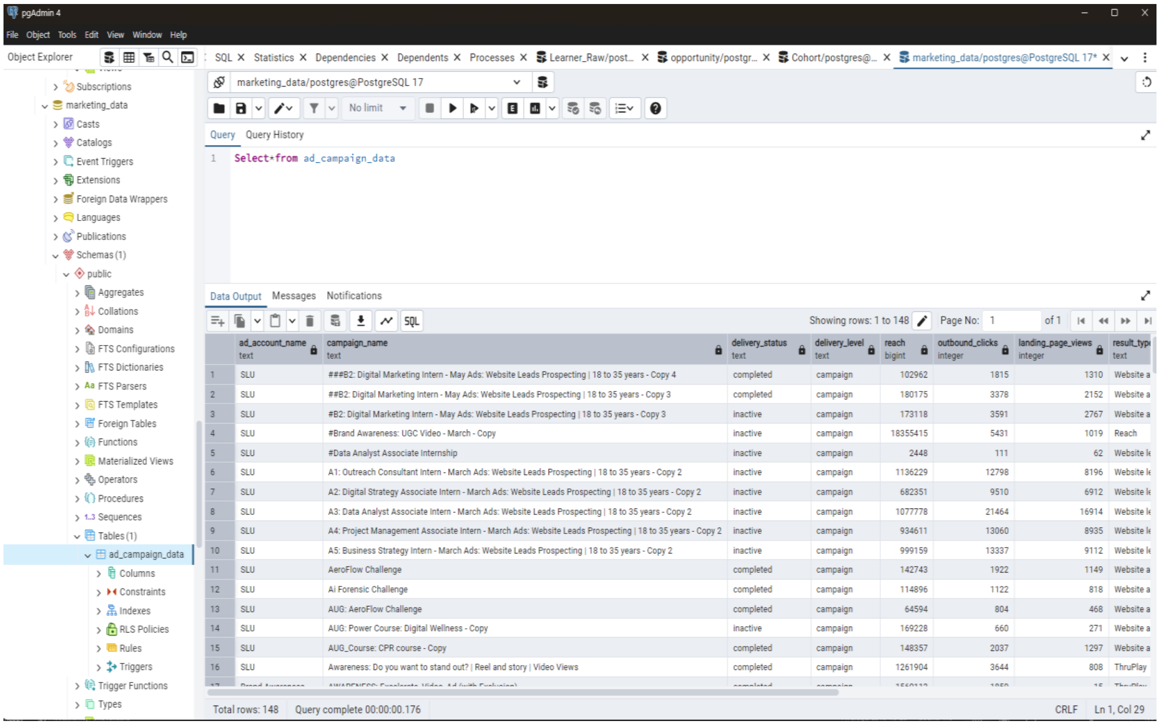
```
CREATE TABLE public.cohort_data (
    cohort_id VARCHAR(20),
    cohort_code VARCHAR(20),
    start_date BIGINT,
    end_date BIGINT,
    size INTEGER
);

select*from Cohort_data
```

Showing rows: 1 to 641 Page No: 1 of 1

	cohort_id character varying (20)	cohort_code character varying (20)	start_date bigint	end_date bigint	size integer
1	Cohort#	B456514	1680500000000	1682960000000	1500
2	Cohort#	B328821	1673850000000	1676910000000	1000
3	Cohort#	B289256	1668400000000	1671080000000	100000
4	Cohort#	B0VCB0F	1663890000000	1663890000000	40
5	Cohort#	B908347	1673240000000	1676310000000	100000
6	Cohort#	B306047	1673240000000	1676310000000	10000
7	Cohort#	B883644	1656650000000	1659240000000	1500
8	Cohort#	B280844	1668400000000	1671080000000	100000
9	Cohort#	B466039	1664770000000	1667450000000	100000
10	Cohort#	B606140	1673240000000	1676310000000	10000
11	Cohort#	B759124	1665370000000	1668530000000	1500
12	Cohort#	B304681	1664770000000	1667450000000	100000
13	Cohort#	B809432	1657890000000	1657890000000	143
14	Cohort#	B495792	1657890000000	1657900000000	1771

Total rows: 641 Query complete 00:00:00.289

4. Image of Marketing_Data successfully imported



marketing_data/postgres@PostgreSQL 17

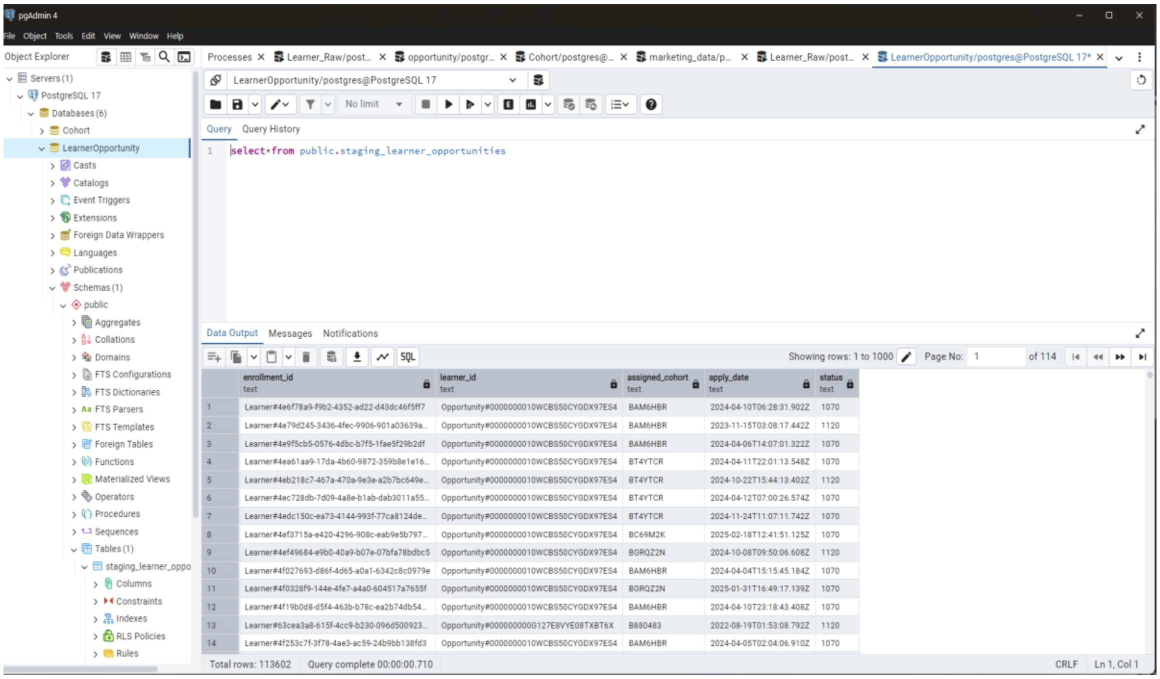
```
Select*from ad_campaign_data
```

Showing rows: 1 to 148 Page No: 1 of 1

	ad_account_name text	campaign_name text	delivery_status text	delivery_level text	reach bigint	outbound_clicks integer	landing_page_views integer	result_type text
1	SLU	###B2: Digital Marketing Intern - May Ads: Website Leads Prospecting | 18 to 35 years - Copy 4	completed	campaign	102962	1815	1310	Website a
2	SLU	##B2: Digital Marketing Intern - May Ads: Website Leads Prospecting | 18 to 35 years - Copy 3	completed	campaign	180175	3378	2152	Website a
3	SLU	#B2: Digital Marketing Intern - May Ads: Website Leads Prospecting | 18 to 35 years - Copy 3	inactive	campaign	173118	3591	2767	Website a
4	SLU	#Brand Awareness: UGC Video - March - Copy	inactive	campaign	18355415	5431	1019	Reach
5	SLU	#Data Analyst Associate Internship	inactive	campaign	2448	111	62	Website le
6	SLU	A1: Outreach Consultant Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	1136229	12798	8196	Website le
7	SLU	A2: Digital Strategy Associate Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	682351	9510	6912	Website le
8	SLU	A3: Data Analyst Associate Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	1077778	21464	16914	Website le
9	SLU	A4: Project Management Associate Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	934611	13060	8935	Website le
10	SLU	A5: Business Strategy Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	999159	13337	9112	Website le
11	SLU	AeroFlow Challenge	completed	campaign	142743	1922	1149	Website a
12	SLU	Ai Forensic Challenge	completed	campaign	114896	1122	818	Website a
13	SLU	AUG: AeroFlow Challenge	completed	campaign	64594	804	468	Website a
14	SLU	AUG: Power Course: Digital Wellness - Copy	inactive	campaign	169228	660	271	Website a
15	SLU	AUG_Course: CPR course - Copy	completed	campaign	148357	2037	1297	Website a
16	SLU	Awareness: Do you want to stand out? | Reel and story | Video Views	completed	campaign	1261904	3644	808	ThruPlay

Total rows: 148 Query complete 00:00:00.176

5. Image of LearnerOpportunity _Raw_Data successfully imported



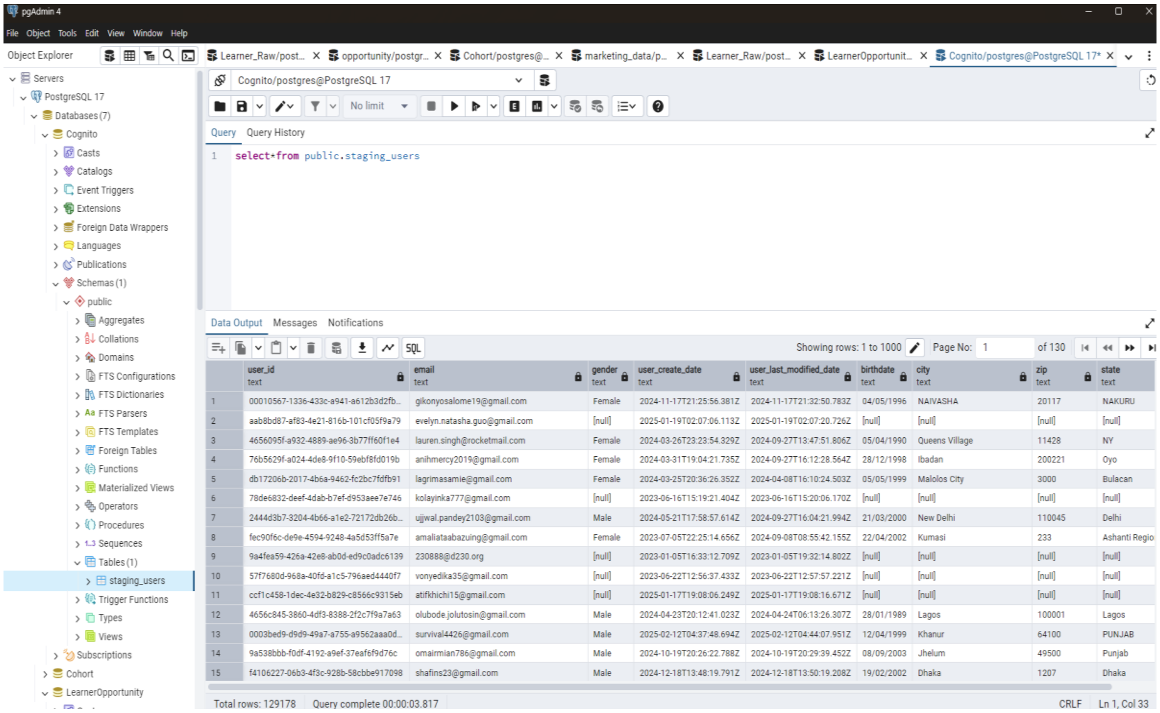
```
select*from public.staging_learner_opportunities
```

Showing rows: 1 to 1000 Page No: 1 of 114

	enrollment_id text	learner_id text	assigned_cohort text	apply_date text	status text
1	Learner#4e6f78a9-f9b2-4352-ad22-d43dc46f5ff7	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2024-04-10T06:28:31.902Z	1070
2	Learner#4e79d245-3436-4fec-9906-901a03639a...	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2023-11-15T03:08:17.442Z	1120
3	Learner#4e9f5cb5-0576-4dbc-b7f5-1fae5f29b2df	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2024-04-06T14:07:01.322Z	1070
4	Learner#4ea61aa9-17da-4b60-9872-359b8e1e16...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	2024-04-11T22:01:13.548Z	1070
5	Learner#4eb218c7-467a-470a-9e3e-a2b7bc649e...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	2024-10-22T15:44:13.402Z	1120
6	Learner#4ec728db-7d09-4a8e-b1ab-dab3011a55...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	2024-04-12T07:00:26.574Z	1070
7	Learner#4edc150c-ea73-4144-993f-77ca8124de...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	2024-11-24T11:07:11.742Z	1070
8	Learner#4ef3715a-e420-4296-908c-eab9e5b797...	Opportunity#0000000010WCBS50CYGDX97ES4	BC69M2K	2025-02-18T12:41:51.125Z	1070
9	Learner#4ef49684-e9b0-40a9-b07e-07bfa78bdbc5	Opportunity#0000000010WCBS50CYGDX97ES4	BGRQZ2N	2024-10-08T09:50:06.608Z	1120
10	Learner#4f027693-d86f-4d65-a0a1-6342c8c0979e	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2024-04-04T15:15:45.184Z	1070
11	Learner#4f0328f9-144e-4fe7-a4a0-604517a7655f	Opportunity#0000000010WCBS50CYGDX97ES4	BGRQZ2N	2025-01-31T16:49:17.139Z	1070
12	Learner#4f19b0d8-d5f4-463b-b78c-ea2b74db54...	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2024-04-10T23:18:43.408Z	1070
13	Learner#63cea3a8-615f-4cc9-b230-096d500923...	Opportunity#000000000G127E8VYE08TXBT6X	B880483	2022-08-19T01:53:08.792Z	1120
14	Learner#4f253c7f-3f78-4ae3-ac59-24b9bb138fd3	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2024-04-05T02:04:06.910Z	1070

Total rows: 113602 Query complete 00:00:00.710 CRLF Ln 1, Col 1

6. Image of Cognito_Data successfully imported



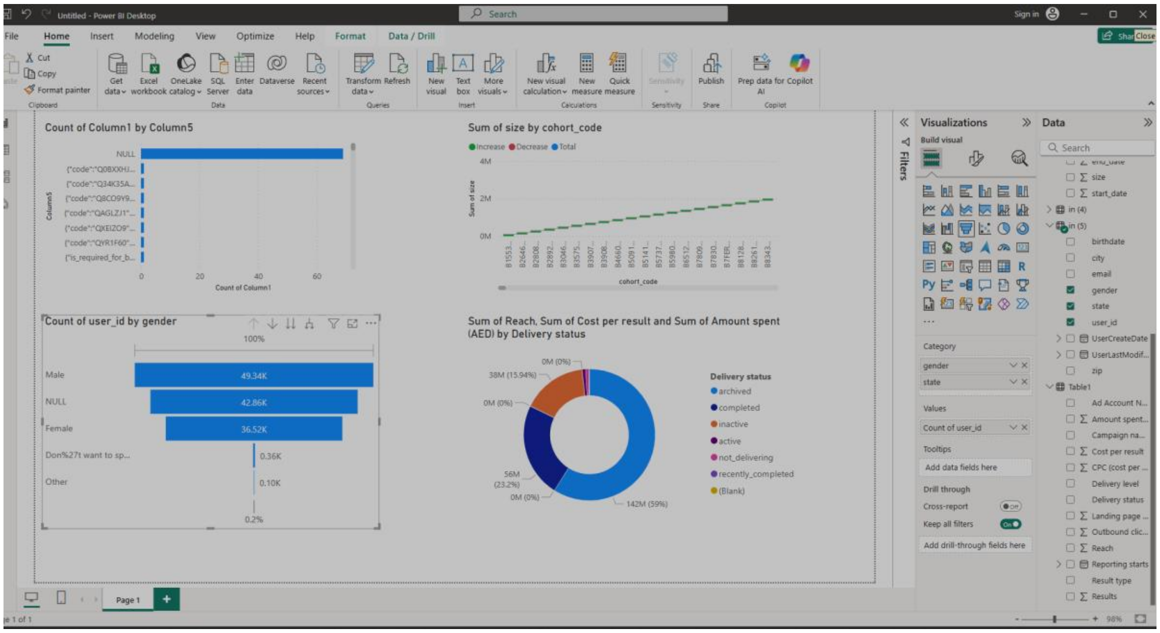
```
select*from public.staging_users
```

Showing rows: 1 to 1000 Page No: 1 of 130

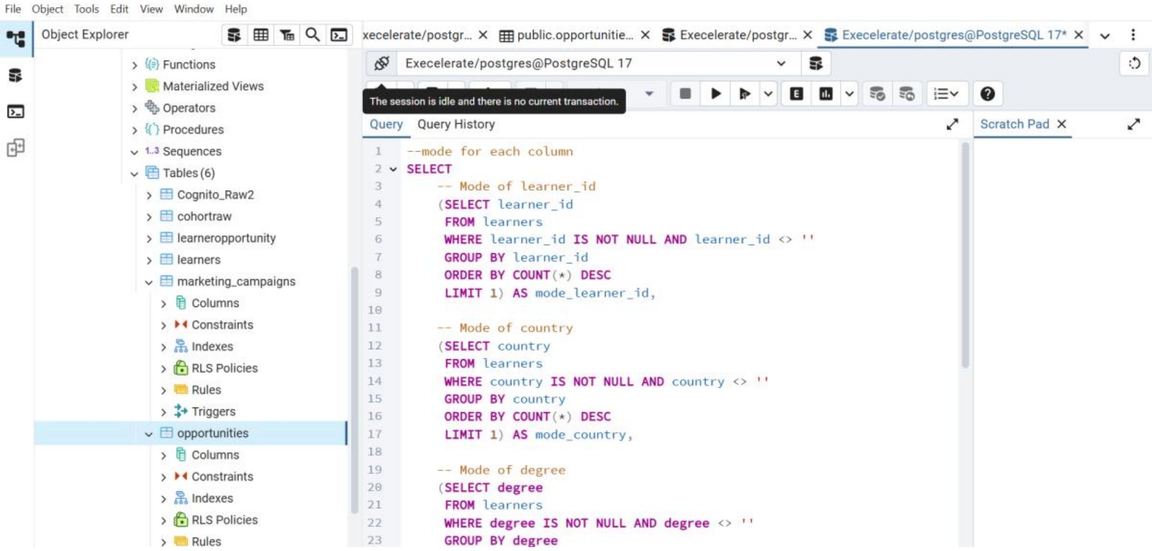
	user_id text	email text	gender text	user_create_date text	user_last_modified_date text	birthdate text	city text	zip text	state text
1	00010567-1336-433c-a941-a612b3d2fb...	gikonyosalome19@gmail.com	Female	2024-11-17T21:25:56.381Z	2024-11-17T21:32:50.783Z	04/05/1996	NAIVASHA	20117	NAKURU
2	aab8bd87-af83-4e21-816b-101cf05f9a79	evelyn.natasha.guo@gmail.com	[null]	2025-01-19T02:07:06.113Z	2025-01-19T02:07:20.726Z	[null]	[null]	[null]	[null]
3	4656095f-a932-4889-ae96-3b77ff60f1e4	lauren.singh@rocketmail.com	Female	2024-03-26T23:23:54.329Z	2024-09-27T13:47:51.806Z	05/04/1990	Queens Village	11428	NY
4	76b5629f-a024-4de8-9f10-59ebf8fd019b	anihmercy2019@gmail.com	Female	2024-03-31T19:04:21.735Z	2024-09-27T16:12:28.564Z	28/12/1998	Ibadan	200221	Oyo
5	db17206b-2017-4b6a-9462-fc2bc7fdfb91	lagrimasamie@gmail.com	Female	2024-03-25T20:36:26.352Z	2024-04-08T16:10:24.503Z	05/05/1999	Malolos City	3000	Bulacan
6	78de6832-deef-4dab-b7ef-d953aee7e746	kolayinka777@gmail.com	[null]	2023-06-16T15:19:21.404Z	2023-06-16T15:20:06.170Z	[null]	[null]	[null]	[null]
7	2444d3b7-3204-4b66-a1e2-72172db26b...	ujjwal.pandey2103@gmail.com	Male	2024-05-21T17:58:57.614Z	2024-09-27T16:04:21.994Z	21/03/2000	New Delhi	110045	Delhi
8	fec90f6c-de9e-4594-9248-4a5d53ff5a7e	amaliataabazuing@gmail.com	Female	2023-07-05T22:25:14.656Z	2024-09-08T08:55:42.155Z	22/04/2002	Kumasi	233	Ashanti Regio
9	9a4fea59-426a-42e8-ab0d-ed9c0adc6139	230888@d230.org	[null]	2023-01-05T16:33:12.709Z	2023-01-05T19:32:14.802Z	[null]	[null]	[null]	[null]
10	57f7680d-968a-40fd-a1c5-796aed4440f7	vonyedika35@gmail.com	[null]	2023-06-22T12:56:37.433Z	2023-06-22T12:57:57.221Z	[null]	[null]	[null]	[null]
11	ccf1c458-1dec-4e32-b829-c8566c9315eb	atifkhichi15@gmail.com	[null]	2025-01-17T19:08:06.249Z	2025-01-17T19:08:16.671Z	[null]	[null]	[null]	[null]
12	4656c845-3860-4df3-8388-2f2c7f9a7a63	olubode.jolutosin@gmail.com	Male	2024-04-23T20:12:41.023Z	2024-04-24T06:13:26.307Z	28/01/1989	Lagos	100001	Lagos
13	0003bed9-d9d9-49a7-a755-a9562aaa0d...	survival4426@gmail.com	Male	2025-02-12T04:37:48.694Z	2025-02-12T04:44:07.951Z	12/04/1999	Khanur	64100	PUNJAB
14	9a538bbb-f0df-4192-a9ef-37eaf6f9d76c	omairmian786@gmail.com	Male	2024-10-19T20:26:22.788Z	2024-10-19T20:29:39.452Z	08/09/2003	Jhelum	49500	Punjab
15	f4106227-06b3-4f3c-928b-58cbbe917098	shafins23@gmail.com	Male	2024-12-18T13:48:19.791Z	2024-12-18T13:50:19.208Z	19/02/2002	Dhaka	1207	Dhaka

Total rows: 129178 Query complete 00:00:03.817 CRLF Ln 1, Col 33

Trend Visuals Using Power BI

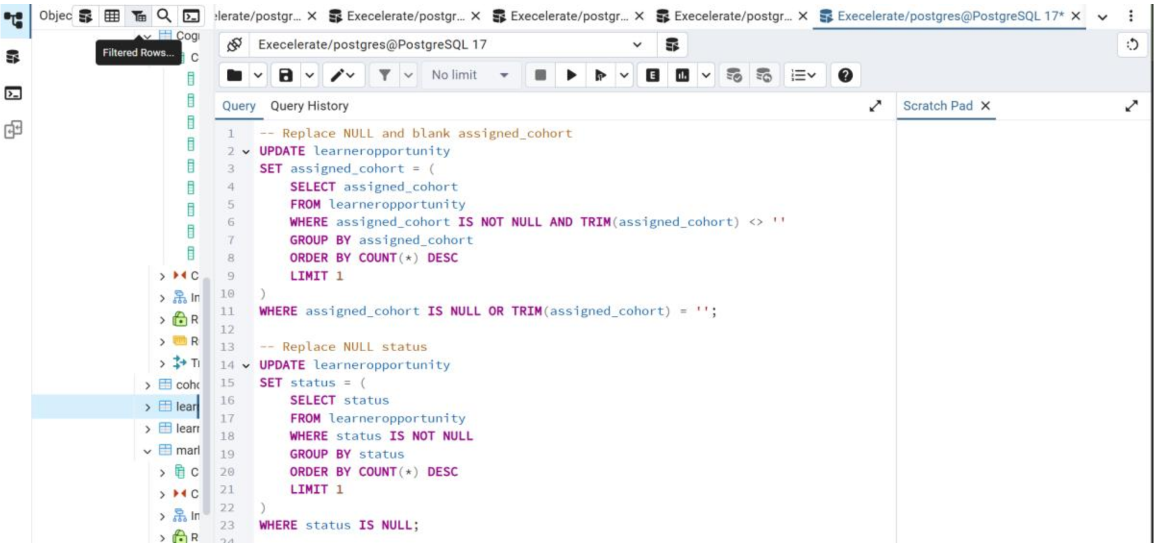


Summarize Data Sample:

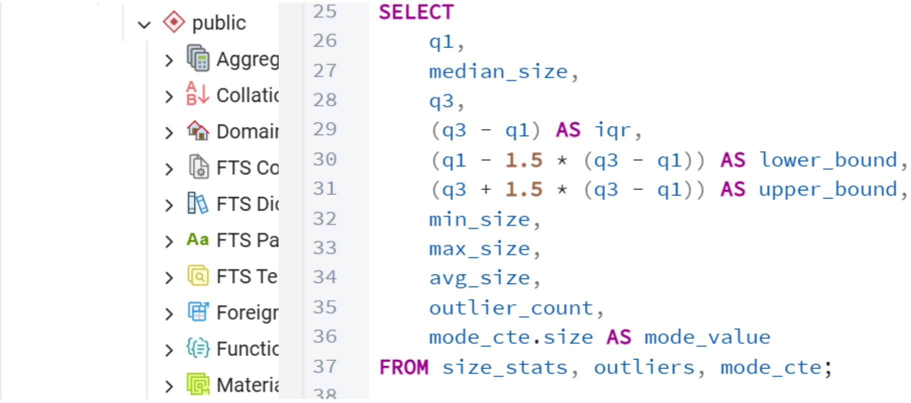


The session is idle and there is no current transaction.

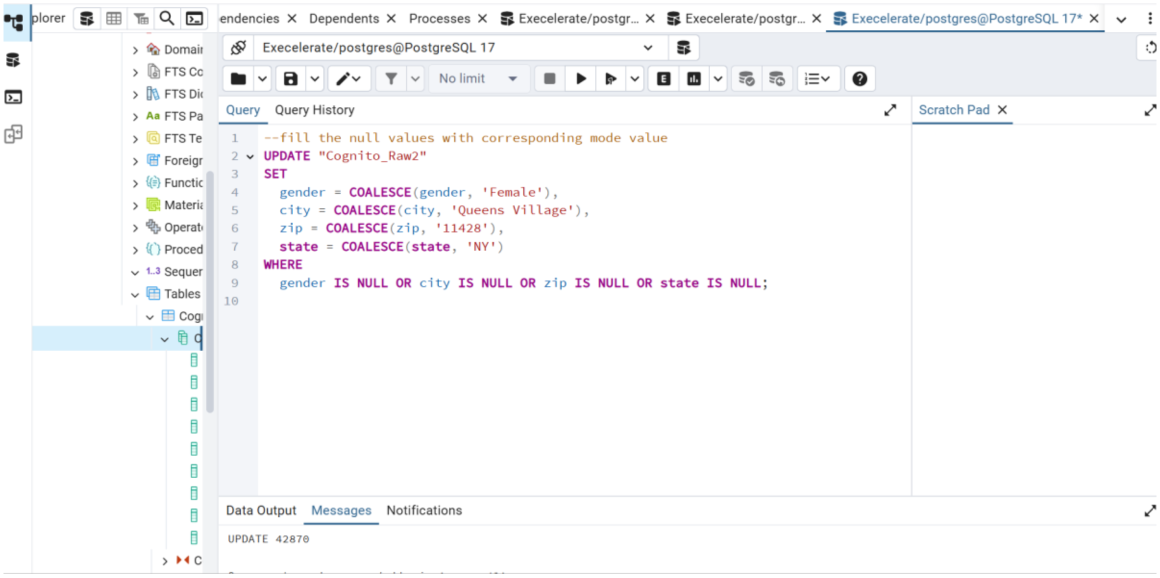
```sql
--mode for each column
SELECT
    -- Mode of learner_id
    (SELECT learner_id
     FROM learners
     WHERE learner_id IS NOT NULL AND learner_id <> ''
     GROUP BY learner_id
     ORDER BY COUNT(*) DESC
     LIMIT 1) AS mode_learner_id,

    -- Mode of country
    (SELECT country
     FROM learners
     WHERE country IS NOT NULL AND country <> ''
     GROUP BY country
     ORDER BY COUNT(*) DESC
     LIMIT 1) AS mode_country,

    -- Mode of degree
    (SELECT degree
     FROM learners
     WHERE degree IS NOT NULL AND degree <> ''
     GROUP BY degree
```


Filtered Rows...
Execelerate/postgres@PostgreSQL 17*
Execelerate/postgres@PostgreSQL 17
No limit
Query
Query History
Scratch Pad

```
-- Replace NULL and blank assigned_cohort
UPDATE learneropportunity
SET assigned_cohort = (
    SELECT assigned_cohort
    FROM learneropportunity
    WHERE assigned_cohort IS NOT NULL AND TRIM(assigned_cohort) <> ''
    GROUP BY assigned_cohort
    ORDER BY COUNT(*) DESC
    LIMIT 1
)
WHERE assigned_cohort IS NULL OR TRIM(assigned_cohort) = '';

-- Replace NULL status
UPDATE learneropportunity
SET status = (
    SELECT status
    FROM learneropportunity
    WHERE status IS NOT NULL
    GROUP BY status
    ORDER BY COUNT(*) DESC
    LIMIT 1
)
WHERE status IS NULL;
```

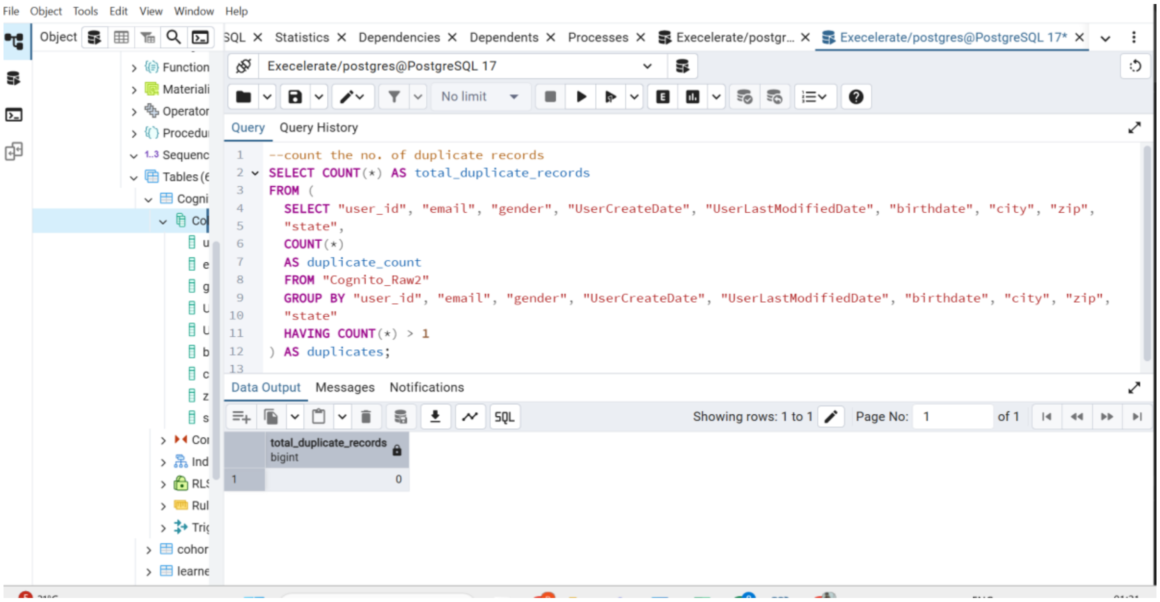


```
SELECT
    q1,
    median_size,
    q3,
    (q3 - q1) AS iqr,
    (q1 - 1.5 * (q3 - q1)) AS lower_bound,
    (q3 + 1.5 * (q3 - q1)) AS upper_bound,
    min_size,
    max_size,
    avg_size,
    outlier_count,
    mode_cte.size AS mode_value
FROM size_stats, outliers, mode_cte;
```



```
--fill the null values with corresponding mode value
UPDATE "Cognito_Raw2"
SET
  gender = COALESCE(gender, 'Female'),
  city = COALESCE(city, 'Queens Village'),
  zip = COALESCE(zip, '11428'),
  state = COALESCE(state, 'NY')
WHERE
  gender IS NULL OR city IS NULL OR zip IS NULL OR state IS NULL;
```

```
UPDATE 42870
```

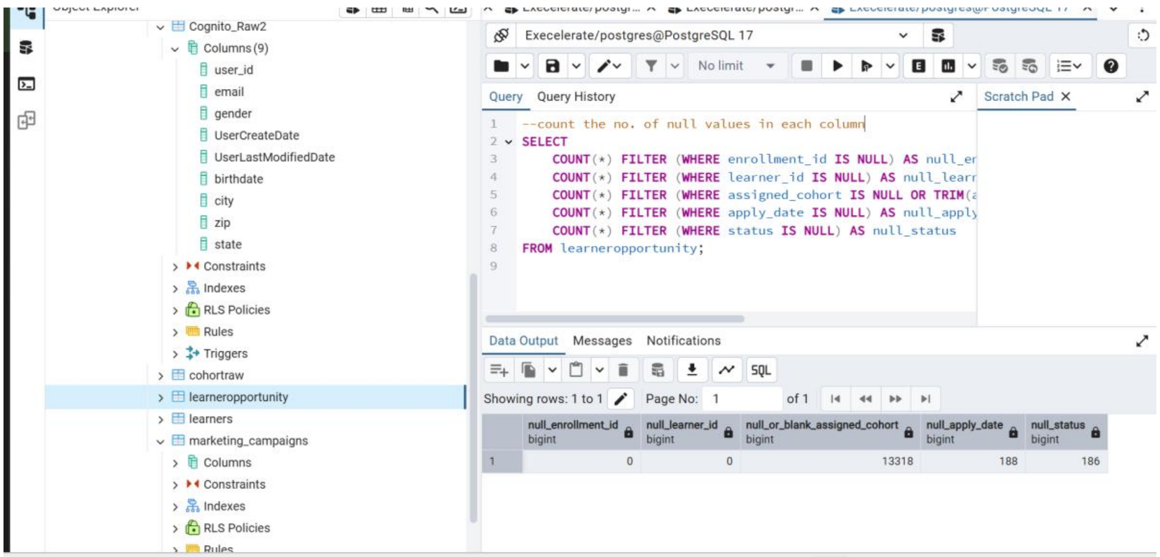

File Object Tools Edit View Window Help
SQL Statistics Dependencies Dependents Processes Execelerate/postgr... Execelerate/postgres@PostgreSQL 17*
Function
Materiali
Operator
Procedu
Sequenc
Tables (6
Cogni
Co
Cor
Ind
RLS
Rul
Trig
cohor
learne
Query
Query History
No limit
Data Output
Messages
Notifications
SQL

Execelerate/postgres@PostgreSQL 17

```
--count the no. of duplicate records
SELECT COUNT(*) AS total_duplicate_records
FROM (
  SELECT "user_id", "email", "gender", "UserCreateDate", "UserLastModifiedDate", "birthdate", "city", "zip",
  "state",
  COUNT(*)
  AS duplicate_count
  FROM "Cognito_Raw2"
  GROUP BY "user_id", "email", "gender", "UserCreateDate", "UserLastModifiedDate", "birthdate", "city", "zip",
  "state"
  HAVING COUNT(*) > 1
) AS duplicates;
```

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

	total_duplicate_records bigint
1	0



Execelerate/postgres@PostgreSQL 17

```
--count the no. of null values in each column
SELECT
    COUNT(*) FILTER (WHERE enrollment_id IS NULL) AS null_er
    COUNT(*) FILTER (WHERE learner_id IS NULL) AS null_learr
    COUNT(*) FILTER (WHERE assigned_cohort IS NULL OR TRIM(a
    COUNT(*) FILTER (WHERE apply_date IS NULL) AS null_apply
    COUNT(*) FILTER (WHERE status IS NULL) AS null_status
FROM learneropportunity;
```

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

	null_enrollment_id bigint	null_learner_id bigint	null_or_blank_assigned_cohort bigint	null_apply_date bigint	null_status bigint
1	0	0	13318	188	186