Data Quality Report of Team 28

Created by-

Kayode Idris Adelakun (ikadelakun@gmail.com)

Most.Sonia Islam (saniaislamsava@gmail.com)

Nahian Tasnim (nahian.tasnim@slu.edu)

Mithun Kumar Dey (mithunkumardey789@gmail.com)

Nirja Patel (nirja2501@gmail.com)

Understanding & Identifying Data Issues

Dataset/Table Structure:

Dataset/Table	Rows	Columns	Data type (Column names)
Learner opportunity data	113602	5	TEXT/object (enrollment_id, learner_id, assigned_cohort, apply_date), DOUBLE PRECISION/float64 (status)
Marketing data	148	14	TEXT /object (Ad Account Name, Campaign name, Delivery status, Delivery level, Result type, Reporting starts), DOUBLE PRECISION /float64 (Reach, Outbound clicks, Landing page views, Results, Cost per result, Amount spent (AED), CPC (cost per link click), rpc)
Opportunity data	187	5	TEXT/object (opportunity_id, opportunity_name, category, opportunity_code, tracking_questions)
Cohort data	639	5	TEXT /object (cohort_id, cohort_code), DOUBLE PRECISION /float64 (start_date, end_date), INTEGER/int64 (size)
User data	129259	5	TEXT /object (learner_id, country, degree, institution, major)
Cognito data	129178	9	TEXT /object (user_id, email, gender, UserCreateDate, UserLastModifiedDate, city, zip, state), TIMESTAMP/datetime64[ns] (birthdate)

Explore Datasets:

Dataset/Table	Key columns that connect datasets	Purpose of dataset
Learner opportunity data	Enrollment_id, learner_id, assigned_cohort, apply_date	Tracking learners' participation in different opportunities, linking user enrollments to programs.
Marketing data	Reporting_starts	Captures advertising performance, including campaign reach, engagement

		metrics, and costs, enabling an analysis of marketing effectiveness on enrollments.
Opportunity data	opportunity_id	Provides information about learning opportunities, including program details, cohort associations, sponsorships, and user participation metrics.
Cohort data	cohort_code	Tracks cohort-based learning programs, including cohort sizes, timelines, and linked opportunities, allowing for participation and completion analysis.
User data	learner_id	Contains user profile details, including demographics, education background, and sign-up timestamps. It helps analyze enrollment trends and user characteristics.
Cognito data	user_id	Contains authentication and profile metadata, including email, gender, and location details.

Data Quality Issues:

Dataset/Table	Columns with missing values	Duplicate records	Inconsistent formats
Learner opportunity data	assigned_cohort, apply_date, status	0	Data type – apply_date, status Inconsistent casing
Marketing data	All columns	5	Data type- Reach, Outbound clicks, Landing page views, Results, Reporting starts Inconsistent casing
Opportunity data	tracking_questions	0	Inconsistent casing
Cohort data	N/A	0	Data type – start_date, end_date Inconsistent casing
User data	country, degree, instituition, major	0	Inconsistent casing
Cognito data	gender, birthdate, city, zip, state	0	Data type- UserCreateDate, UserLastModifiedDate, birthdate, zip Inconsistent casing

Building the Master Table & ETL Process

Plan the master table:

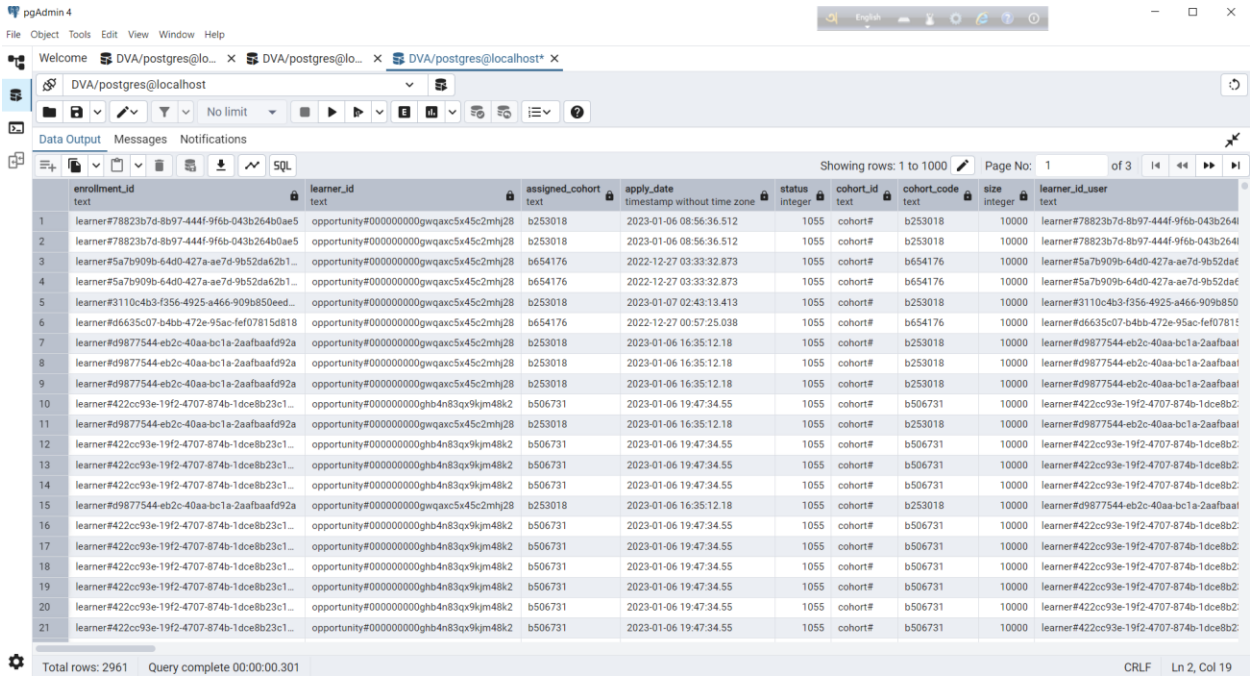
Datasets/Table	Key columns	Primary key	Foreign keys
Learner opportunity data	enrollment_id, learner_id, assigned_cohort	Composite primary key (enrollment_id + learner_id)	enrollment_id, learner_id, assigned_cohort
Marketing data	ad Account Name, campaign name	N/A	reporting_starts
Opportunity data	opportunity_id, opportunity_code	opportunity_id	opportunity-id
Cohort data	cohort_code	cohort_id	cohort_code
User data	learner_id, country	learner_id	learner_id
Cognito data	user_id, email	user_id	user_id

Validation & Refinement

Data Quality Checks and Refinement:

Shape of master table		2961 rows, 42 columns
Record count validation (Rows)	Raw datasets	486428 (Total record for all 6 datasets)
	Master table	2961
	Root cause	Having missing and duplicate values in raw datasets.
Duplicate checks		No duplicate records
Missing Data		No missing data
Foreign key integrity		All linked record exist in the given datasets (changed the learner_id column name of user dataset to learner_id_user as there is another learner_id column in learner opportunity dataset
Data type verification		All columns are in correct format

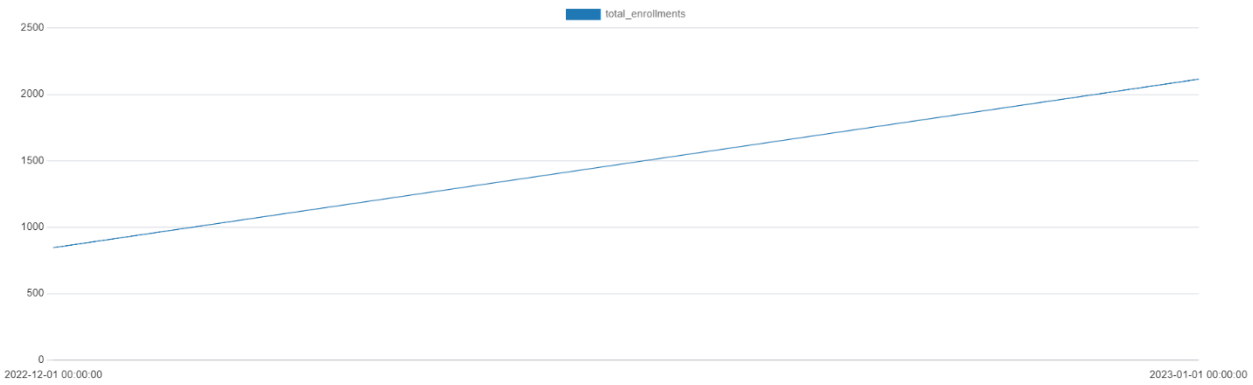
Master Table:



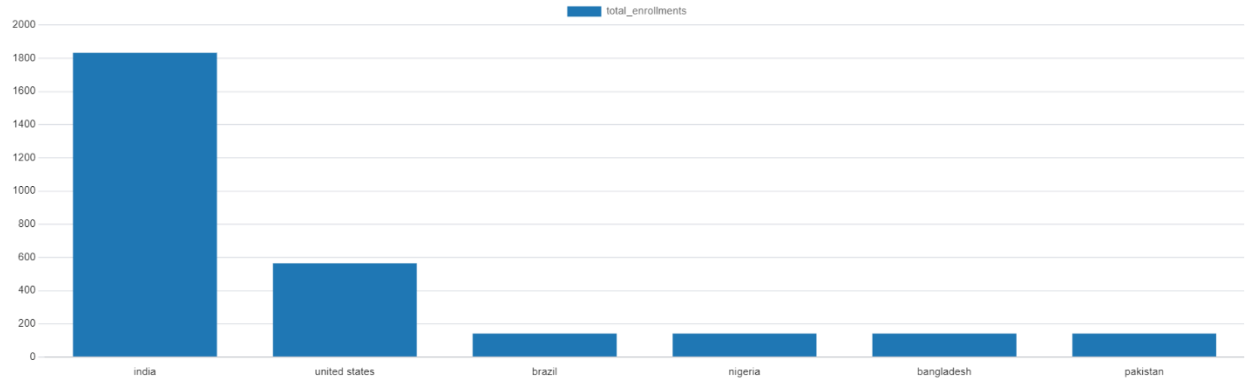
	enrollment_id text	learner_id text	assigned_cohort text	apply_date timestamp without time zone	status integer	cohort_id text	cohort_code text	size integer	learner_id_user text
1	learner#78823b7d-8b97-444f-9f6b-043b264b0ae5	opportunity#000000000gwqaxc5x45c2mhj28	b253018	2023-01-06 08:56:36.512	1055	cohort#	b253018	10000	learner#78823b7d-8b97-444f-9f6b-043b264l
2	learner#78823b7d-8b97-444f-9f6b-043b264b0ae5	opportunity#000000000gwqaxc5x45c2mhj28	b253018	2023-01-06 08:56:36.512	1055	cohort#	b253018	10000	learner#78823b7d-8b97-444f-9f6b-043b264l
3	learner#5a7b909b-64d0-427a-ae7d-9b52da62b1...	opportunity#000000000gwqaxc5x45c2mhj28	b654176	2022-12-27 03:33:32.873	1055	cohort#	b654176	10000	learner#5a7b909b-64d0-427a-ae7d-9b52daE
4	learner#5a7b909b-64d0-427a-ae7d-9b52da62b1...	opportunity#000000000gwqaxc5x45c2mhj28	b654176	2022-12-27 03:33:32.873	1055	cohort#	b654176	10000	learner#5a7b909b-64d0-427a-ae7d-9b52daE
5	learner#3110c4b3-f356-4925-a466-909b850eed...	opportunity#000000000gwqaxc5x45c2mhj28	b253018	2023-01-07 02:43:13.413	1055	cohort#	b253018	10000	learner#3110c4b3-f356-4925-a466-909b850
6	learner#d6635c07-b4bb-472e-95ac-fef07815d818	opportunity#000000000gwqaxc5x45c2mhj28	b654176	2022-12-27 00:57:25.038	1055	cohort#	b654176	10000	learner#d6635c07-b4bb-472e-95ac-fef07815
7	learner#d9877544-eb2c-40aa-bc1a-2aafbaafd92a	opportunity#000000000gwqaxc5x45c2mhj28	b253018	2023-01-06 16:35:12.18	1055	cohort#	b253018	10000	learner#d9877544-eb2c-40aa-bc1a-2aafbaal
8	learner#d9877544-eb2c-40aa-bc1a-2aafbaafd92a	opportunity#000000000gwqaxc5x45c2mhj28	b253018	2023-01-06 16:35:12.18	1055	cohort#	b253018	10000	learner#d9877544-eb2c-40aa-bc1a-2aafbaal
9	learner#d9877544-eb2c-40aa-bc1a-2aafbaafd92a	opportunity#000000000gwqaxc5x45c2mhj28	b253018	2023-01-06 16:35:12.18	1055	cohort#	b253018	10000	learner#d9877544-eb2c-40aa-bc1a-2aafbaal
10	learner#422cc93e-19f2-4707-874b-1dce8b23c1...	opportunity#000000000ghb4n83qx9kjm48k2	b506731	2023-01-06 19:47:34.55	1055	cohort#	b506731	10000	learner#422cc93e-19f2-4707-874b-1dce8b2
11	learner#d9877544-eb2c-40aa-bc1a-2aafbaafd92a	opportunity#000000000gwqaxc5x45c2mhj28	b253018	2023-01-06 16:35:12.18	1055	cohort#	b253018	10000	learner#d9877544-eb2c-40aa-bc1a-2aafbaal
12	learner#422cc93e-19f2-4707-874b-1dce8b23c1...	opportunity#000000000ghb4n83qx9kjm48k2	b506731	2023-01-06 19:47:34.55	1055	cohort#	b506731	10000	learner#422cc93e-19f2-4707-874b-1dce8b2
13	learner#422cc93e-19f2-4707-874b-1dce8b23c1...	opportunity#000000000ghb4n83qx9kjm48k2	b506731	2023-01-06 19:47:34.55	1055	cohort#	b506731	10000	learner#422cc93e-19f2-4707-874b-1dce8b2
14	learner#422cc93e-19f2-4707-874b-1dce8b23c1...	opportunity#000000000ghb4n83qx9kjm48k2	b506731	2023-01-06 19:47:34.55	1055	cohort#	b506731	10000	learner#422cc93e-19f2-4707-874b-1dce8b2
15	learner#d9877544-eb2c-40aa-bc1a-2aafbaafd92a	opportunity#000000000gwqaxc5x45c2mhj28	b253018	2023-01-06 16:35:12.18	1055	cohort#	b253018	10000	learner#d9877544-eb2c-40aa-bc1a-2aafbaal
16	learner#422cc93e-19f2-4707-874b-1dce8b23c1...	opportunity#000000000ghb4n83qx9kjm48k2	b506731	2023-01-06 19:47:34.55	1055	cohort#	b506731	10000	learner#422cc93e-19f2-4707-874b-1dce8b2
17	learner#422cc93e-19f2-4707-874b-1dce8b23c1...	opportunity#000000000ghb4n83qx9kjm48k2	b506731	2023-01-06 19:47:34.55	1055	cohort#	b506731	10000	learner#422cc93e-19f2-4707-874b-1dce8b2
18	learner#422cc93e-19f2-4707-874b-1dce8b23c1...	opportunity#000000000ghb4n83qx9kjm48k2	b506731	2023-01-06 19:47:34.55	1055	cohort#	b506731	10000	learner#422cc93e-19f2-4707-874b-1dce8b2
19	learner#422cc93e-19f2-4707-874b-1dce8b23c1...	opportunity#000000000ghb4n83qx9kjm48k2	b506731	2023-01-06 19:47:34.55	1055	cohort#	b506731	10000	learner#422cc93e-19f2-4707-874b-1dce8b2
20	learner#422cc93e-19f2-4707-874b-1dce8b23c1...	opportunity#000000000ghb4n83qx9kjm48k2	b506731	2023-01-06 19:47:34.55	1055	cohort#	b506731	10000	learner#422cc93e-19f2-4707-874b-1dce8b2
21	learner#422cc93e-19f2-4707-874b-1dce8b23c1...	opportunity#000000000ghb4n83qx9kjm48k2	b506731	2023-01-06 19:47:34.55	1055	cohort#	b506731	10000	learner#422cc93e-19f2-4707-874b-1dce8b2

Visualizations

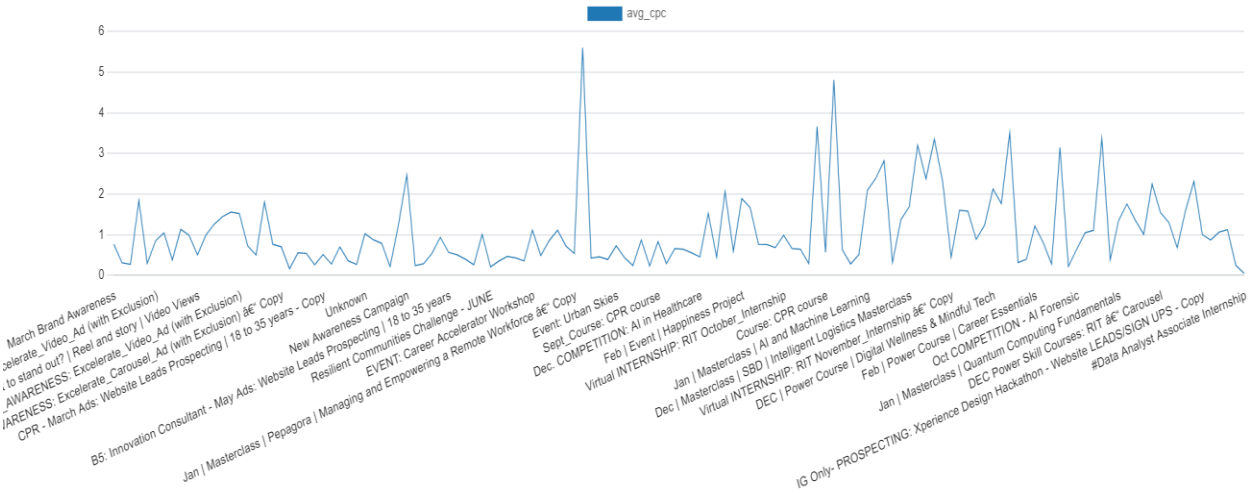
Enrollment Trend Over Time:



Top Countries by Enrollment:



Average CPC per campaign:



Enrollment by Gender: