EDA Report

Data Sets:

Cognito_Raw2: A user dataset with columns: userid, email, gender, usercreateddate, userlastmodifieddate, birthdate, city, zip, state.

CohortRaw: A cohort dataset with columns: cohortid, cohortcode, startdate, enddate, size

Learner_Raw: A student dataset with columns: learnerid, country, degree, institution, major

LearnerOpportunity_Raw: A course dataset with columns: enrollmentid, learnerid, asignedcohort, applydate, status

MarketingCampaignDataAllAccounts: A campaign dataset with columns: accountname, campaignname, deliverystatus, deliverylevel, reach, outboundclicks, outboundtype, resulttype, results, costperresult, amountspent, cpc, dates

OpportunityRaw: A job dataset with columns: opportunityid, opportunityname, category, opportunitycode, trackingquestions

Summary Statistics:

1. Cognito_Raw2:

Key Statistics

Gender Distribution (Non-NULL values)

- Male: 263 (49.3%)
- Female: 263 (49.3%)
- Other: 7 (1.3%)

Temporal Analysis

- Account creation dates range from 2023-01-05 to 2025-02-24
- Most recent account modification: 2025-02-24
- The dataset shows consistent user acquisition over time with spikes in certain periods

Geographic Distribution

Top 5 states/regions:

- Lagos, Nigeria: 28 records
- Maharashtra, India: 15 records
- Dhaka, Bangladesh: 12 records
- Nairobi, Kenya: 11 records
- Andhra Pradesh, India: 10 records

The data shows a strong representation from South Asia and Africa, with users from 45+ different countries.

2. LearnerOpportunity_Raw:

Variable Analysis

1. enrollment_id

- **Type**: Unique identifier (string)
- **Format**: "Learner#UUID"
- **Unique values**: All appear to be unique (primary key)

2. learner_id

- **Type**: Identifier (string)
- **Format**: "Opportunity#ID"
- **Unique values**: Multiple learners share the same opportunity ID
- **Most common opportunities**:
- "0000000010WCBS50CYGDX97ES4" (majority)
- "000000000G127E8VYE08TXBT6X"
- "000000000G4AM4J9NBMPK3TJH6"
- "0000000010GJ8R10FT5FETZ366"
- "000000000GRABD28CXVVEYEX21"

3. assigned_cohort

- **Type**: Categorical
- **Unique values**: 7
- **Most common cohorts**:

- BAM6HBR (most frequent)
- BGRQZ2N
- BT4YTCR
- BC69M2K
- B880483
- BOQ24ZG
- B251425
- BC9A1OB
- NULL (1 record)

4. apply_date

- **Type**: DateTime (ISO format)
- **Range**: 2022-08-19 to 2025-02-23
- **Distribution**:
- Most applications are clustered around April 2024
- Some historical applications from 2022-2023
- Future-dated applications extend into 2025

5. status

- **Type**: Numeric code
- **Unique values**: 10 distinct codes
- **Most common statuses**:
- 1070 (most frequent)
- 1120
- 1080
- 1110
- 1055
- 1030

Key Insights

- **Temporal Patterns**:

- There's a significant spike in applications in April 2024
- Some applications are dated in the future (up to February 2025)
- **Cohort Distribution**:
- BAM6HBR is the most frequently assigned cohort
- Most records are distributed among 4 main cohorts (BAM6HBR, BGRQZ2N, BT4YTCR, BC69M2K)
- **Status Codes**:
- 1070 is by far the most common status
- The meaning of these codes would require additional documentation
- **Data Quality**:
- One record has a NULL value for assigned_cohort
- All other records have complete data

3. Opportunity_Raw:
Distribution of Opportunities by Category:

Category	Count	Percentage
Internship	70	35%
Competition	55	27.5%
Course	42	21%
Career	20	10%
Masterclass	7	3.5%

Missing and Duplicate Values:

There are many missing and duplicate values in the data sets but the major and most populated ones are:

Birthdate column from Cognito_Raw2

Degree, Institution, Major columns from Learner_Raw

Tracking_questions column from Opportunity_Raw

Outliers and Anomalies:

1. Cognito_Raw2:

Data Quality Assessment

Missing Values

- gender: 155 NULL values (22.5% of records)
- birth_date: 155 NULL values (22.5% of records)
- city: 155 NULL values (22.5% of records)
- zip: 155 NULL values (22.5% of records)
- state: 155 NULL values (22.5% of records)

Note: The same 155 records have NULL values across all these fields, suggesting these may be incomplete registrations.

Data Trends:

The **CohortRaw, Learner_Raw** and **LearnerOpportunity_Raw** have a very strong relationship in between them as the primary key of **Learner_Raw** (learnerid) is a foreign key in **LearnerOpportunity_Raw** and the primary key of **CohortRaw** (cohortid) is also a foreign key in **LearnerOpportunity_Raw** so this is a relationship we founded pretty obvious in the data sets.