

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, assignedcohort, 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. CohortRaw

Columns: cohort_id, cohort_code, start_date, end_date, size

Key Observations:

Size:

Mean: ~9,000

Min: 3

Max: 100,000

2. Marketing Campaign Data All Accounts (2023-2024)

Columns: Ad Account, Campaign Name, Delivery Status, Reach, Clicks, Results, Cost, etc.

Key Metrics:

Reach:

Mean: ~1.5M

Min: 0

Max: 141M

Cost per Result:

Mean: ~3.5 AED

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. CohortRaw

Outliers:

Very small cohorts (e.g., size = 3, 10, 11) may represent test groups or errors.

Very large cohorts (e.g., size = 100,000) could be bulk entries or data errors.

Date Ranges:

Most cohorts span days to months, but some have extremely short durations (e.g., 1 hour) or long durations (e.g., years).

Anomalies:

Cohorts with identical start/end dates (e.g., B0VCB0F: 1-hour duration).

Future-dated cohorts (e.g., BSGY1WA starting in 2025).

2. Marketing Campaign Data All Accounts (2023-2024)

Outliers:

Extremely high

Extremely low

Results:

Some campaigns show 0 results despite high spend

Delivery Status:

71% completed, 20% inactive, and some "not_delivering" campaigns with spend

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.