# Data Familiarization Report

## DATA VISUALIZATION ASSOCIATE EARLY INTERNSHIP

Exploratory Data Analysis Report

Internship Project: Week 1 Deliverable

Title: Data Understanding, EDA, and Cleaning

Organization: Excelerate


| Sl.no | Name(email) |
|---|---|
| 1. | Neha Sunil(nehasunil588@gmail.com) |
| 2. | Sujay Kumar(sujaykumar94318@gmail.com) |
| 3. | Jayasree Chakraborty (jayasree24@gmail.com) |
| 4. | Sitesh Gupta (guptasistesh03@gmail.com) |

## Report Summary:


This report presents the **Week 1 deliverables** of the internship project, which includes the successful installation and configuration of the PostgreSQL database environment. All required CSV datasets were carefully imported and structured, ensuring proper handling of headers, missing columns, and formatting issues. Data tables were verified for schema integrity, and necessary columns were cleaned or created (e.g., clean_cohort_id) to support further analysis.

The focus of this week was performing Exploratory Data Analysis (EDA) across six datasets related to learners, opportunities, cohort mapping, and marketing campaigns. Detailed SQL queries were used to assess missing values, identify duplicates, extract key statistics, and detect logical inconsistencies such as invalid date sequences or unnormalized formats. Data quality issues, like inconsistent country names, blank majors, and duplicate email entries, were cleaned or flagged for correction.

Initial insights reveal strong engagement trends during late 2022, with status 1070 dominating learner status logs, internships being the most common opportunity type, and cybersecurity being a top area of interest. The marketing dataset highlighted the most clicked campaigns and spend patterns, offering a baseline understanding of outreach effectiveness. These foundational insights and clean datasets prepare the ground for deeper analysis and modelling in the upcoming weeks.

# 1. Introduction

This report presents the **Exploratory Data Analysis (EDA)** for six datasets provided as part of the learning platform data ecosystem. The objective is to understand the structure, quality, and contents of each dataset in preparation for data cleaning and further analytical processes such as dashboarding and modelling.

**EDA helps:**

- **Uncover data quality issues (missing values, duplicates, etc.)**
- **Understand data distributions, relationships, and trends**
- **Prepare data for meaningful visualizations and insights**

# 2. Dataset Overview

| Dataset Name | File Name | Description |
|---|---|---|
| User Data | user_data.csv | Contains user profiles including demographics, education, and sign-up timestamps. Useful for analysing user characteristics and enrollment trends. |
| Opportunity Data | opp_data.csv | Includes program details, cohort associations, sponsorships, and user participation metrics. Useful for studying program performance and reach. |
| Cohort Data | cohort_data.csv | Tracks cohort-based learning groups with details like cohort sizes, dates, and linked opportunities. Enables cohort-level analysis. |
| Marketing Data | marketing_data.csv | Captures campaign-level performance, engagement, and costs. Useful for assessing marketing effectiveness. |
| Learner Opportunity Data | learner_opportunity_raw.csv | Maps learners to the opportunities they enrolled in. Helps in tracking participation and engagement per program. |
| Cognito Data | cognito_raw.csv | Contains metadata like email, gender, and location. Supports user segmentation and profile enrichment. |

# 3. Analysis per Dataset

This section summarizes key findings from the Exploratory Data Analysis (EDA) and outlines the cleaning steps performed for each dataset. It includes the structure of each dataset, missing values, duplicates, inconsistencies, and corrective actions taken.

### 1. Key Observations:

- Missing values found in major and degree
- Inconsistent formatting in country (e.g., "India", "India", " INDIA ")
- No duplicate user_id values

### 2. Data Cleaning Performed:

- Filled missing major with "Unknown"
- Standardized country values using UPDATE "Learner Raw" SET country = INITCAP (TRIM (country));
- Converted signup date to correct DATE format

## 3. Dataset Analysis

To understand the structure and volume of the dataset, we began with a few basic queries:

### !. First 10 data set structure

SELECT * FROM "Learner_Raw" LIMIT 10;

| | learner_id text | country text | degree text | institution text | major text |
|---|---|---|---|---|---|
| 1 | Learner#206dfcde-d5a8-40aa-875e-ece317d801c9 | India | Null | NULL | NULL |
| 2 | Learner#218ca847-eb61-4524-8824-dbfc2fe64c3d | Philippines | Null | NULL | NULL |
| 3 | Learner#22f2aab4-92fd-40fa-ab7f-730c18de877d | United States | Null | NULL | NULL |
| 4 | Learner#2356ae15-e7f5-4213-b61f-d1a9b95fe011 | Egypt | Null | NULL | NULL |
| 5 | Learner#2376bb1b-2514-4599-843a-16b7d34bbd... | Kenya | Null | NULL | NULL |
| 6 | Learner#19942037-f2c4-4946-8b9d-30bb7c6e1d8c | Kenya | Null | NULL | NULL |
| 7 | Learner#25717f27-93bd-4218-9063-94eadf672aac | India | Graduate Student | krishna university | MCA |
| 8 | Learner#1a87ee0d-545f-4935-971c-79f6c683be6f | Bangladesh | Null | NULL | NULL |
| 9 | Learner#27ad8d48-2034-47c3-924e-1d4199033c... | Nigeria | Null | NULL | NULL |
| 10 | Learner#1b126798-441e-49b0-940f-aa5329882809 | Egypt | Null | NULL | NULL |

## 2.Total Count

SELECT COUNT(*) FROM "Learner Raw";

| Count |
|---|
| 129260 |

## 3. Datatype

SELECT column_name, data_type FROM information_schema.columns WHERE table_name = 'Learner_Raw';

Output:

| Column Name | Data Type |
|---|---|
| user_id | INTEGER |
| country | TEXT |
| degree | TEXT |
| major | TEXT |
| signup_date | DATE |

## 4. Selecting Institution Count

SELECT institution, COUNT(*)

FROM "Learner_Raw"

GROUP BY institution

HAVING COUNT(*)>1;

**Output:**

| | country text | count bigint |
|---|---|---|
| 1 | Afghanistan | 176 |
| 2 | Aland Islands | 65 |
| 3 | Albania | 49 |
| 4 | Algeria | 58 |
| 5 | American Samoa | 50 |
| 6 | Andorra | 2 |
| 7 | Angola | 11 |
| 8 | Antarctica | 6 |
| 9 | Argentina | 6 |
| 10 | Armenia | 5 |
| 11 | Aruba | 2 |
| 12 | Australia | 45 |
| 13 | Austria | 4 |
| 14 | Azerbaijan | 7 |
| 15 | Bahamas | 2 |
| 16 | Bahrain | 2 |
| 17 | Bangladesh | 1845 |
| 18 | Belgium | 7 |
| 19 | Belize | 4 |
| 20 | Benin | 13 |
| 21 | Bhutan | 9 |

### 5. Standardize country Field Format

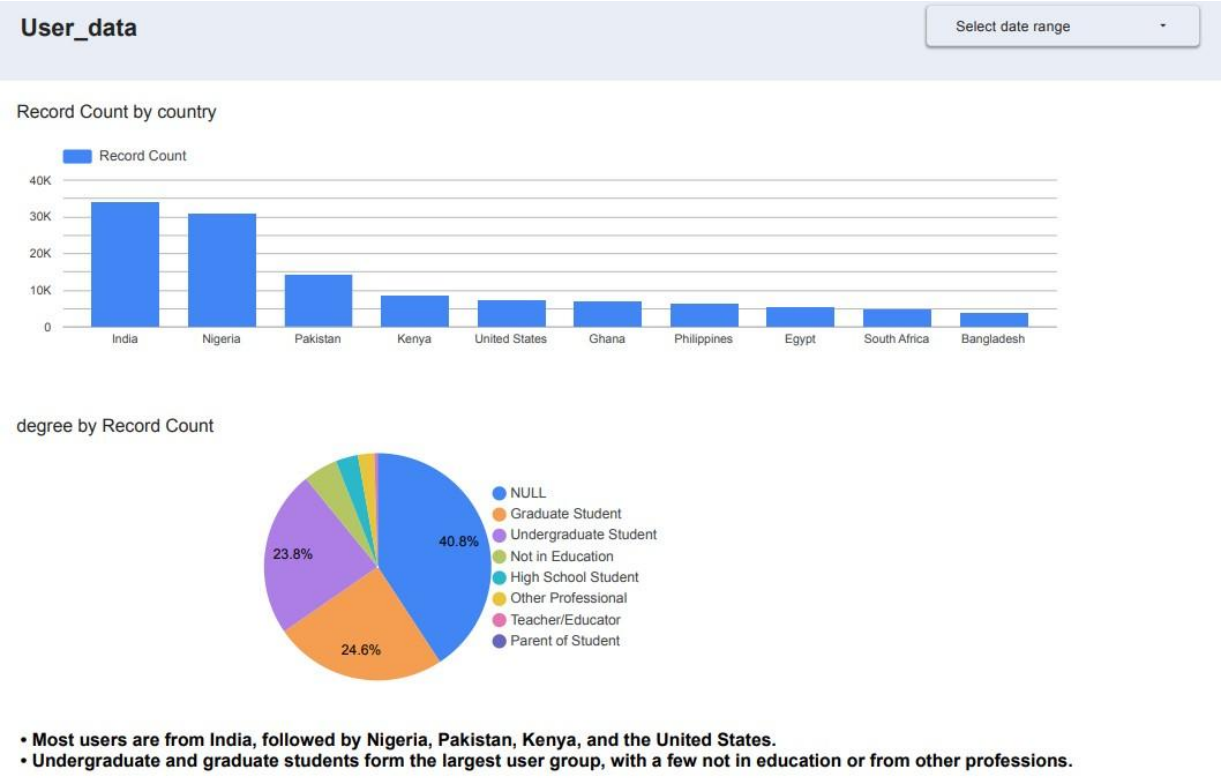
```
UPDATE "Learner_Raw"

    SET country = INITCAP(TRIM(country));
```

**Output:** UPDATE 129260

## Chart Description for User Data Dashboard

#### 1. Bar Chart: Record Count by Country

This bar chart visualizes the number of users by country. The majority of users are from India, followed by Nigeria, Pakistan, Kenya, and the United States. India and Nigeria alone account for a significant portion of the user base, suggesting a strong concentration of learners from South Asia and Africa. Countries like Ghana, the Philippines, Egypt, South Africa, and Bangladesh also contribute to the user pool but in smaller numbers.




## Conclusion:

The data cleaning process for the Learner_Raw dataset successfully improved data quality, consistency, and readiness for analysis. Key issues identified included missing values in the major column, inconsistent casing and spacing in the country and degree fields, and potential duplicate entries based on learner_id.

# B. Opportunity Data Analysis

**Introduction:** The Opportunity_Raw dataset contains metadata about various learning or training opportunities available to users. The primary objective of this exploratory data analysis (EDA) was to examine the completeness, consistency, and distribution of the data across key fields such as opportunity_id, opportunity_name, category, opportunity_code, and tracking_questions.

We began by previewing the dataset and analyzing the total number of records. Key checks included identifying missing values, assessing data uniqueness, and understanding how opportunities are categorized. The analysis also focused on flagging incomplete or missing tracking questions and spotting duplicate opportunity_code values, which could indicate redundant entries.

**1. Data Preview**

```
SELECT opportuniy_id,opportuniy_name,category, opportunity_code,
tracking_questions

FROM "Opportunity_Raw"

LIMIT 10;
```

**Output:**

| | opportunity_id text | opportunity_name text | category text | opportunity_code text | tracking_questions text |
|---|---|---|---|---|---|
| 1 | Opportunity#000000000G8BW90E86ARRKM3... | Cybersecurity: Defensive Hacking | Internship | I155449 | NULL |
| 2 | Opportunity#0000000010RVQ9RQKZ2P2XSZ... | Graphic Design Intern | Career | AVCVHR9 | NULL |
| 3 | Opportunity#000000123GBZ5VRTC3YS9T716N | Data Visualization | Internship | I866009 | NULL |
| 4 | Opportunity#0000000010SAZXDAE05AN2CG... | Project Management Associate Early Internship | Internship | IP5EYAL | {code:Q34K35A,is_requi |
| 5 | Opportunity#0000000010AWJ1XABSV8Y81F... | Business Development Virtual Internship | Internship | I2KYO99 | {serial_number:1,is_requ |
| 6 | Opportunity#0000000010NQ273YXVEPW10... | Mastering Cybersecurity: Safeguarding Confidentiality and Integrity | Masterclass | S667YXU | {serial_number:1,is_requ |
| 7 | Opportunity#0000000010WCBS50CYGDX97E... | CPR/AED Certification | Course | USNZAI6 | {is_required_for_badge_ |
| 8 | Opportunity#000000000G8JG2FEA12SVNXX... | Esports and Game Design | Internship | I860340 | NULL |
| 9 | Opportunity#000000000G95BD07NB0181K0... | Data Visualization | Internship | I660879 | NULL |
| 10 | Opportunity#000000000GBZ5VRTC3YS9T716N | Data Visualization | Internship | I755008 | NULL |

**2. Missing Values Check**

```
SELECT
  COUNT(*) AS total_rows,
  COUNT(opportuniy_id) AS non_null_opportuniy_id,
  COUNT(opportuniy_name) AS non_null_opportuniy_name,
  COUNT(category) AS non_null_category,
  COUNT(opportunity_code) AS non_null_opportunity_code,
  COUNT(tracking_questions) AS non_null_tracking_questions
FROM "Opportunity_Raw";
```

**Output:**

| | total_rows bigint | non_null_opportuniy_id bigint | non_null_opportuniy_name bigint | non_null_category bigint | non_null_opportunity_code bigint | non_null_tracking_questions bigint |
|---|---|---|---|---|---|---|
| 1 | 374 | 374 | 374 | 374 | 374 | 374 |

**3. Uniqueness/Redundancy Check**

```
SELECT
  COUNT(DISTINCT opportuniy_id) AS unique_ids,
  COUNT(DISTINCT opportuniy_name) AS unique_names,
  COUNT(DISTINCT category) AS unique_categories,
  COUNT(DISTINCT opportunity_code) AS unique_codes
FROM "Opportunity_Raw";
```

**Output:**

| | unique_ids bigint | unique_names bigint | unique_categories bigint | unique_codes bigint |
|---|---|---|---|---|
| 1 | 187 | 170 | 7 | 187 |

### 4. Category Distribution

SELECT category, COUNT(*) AS count

FROM "Opportunity_Raw"

GROUP BY category

ORDER BY count DESC;

**Output:**

| | category text | count bigint |
|---|---|---|
| 1 | Internship | 86 |
| 2 | Event | 82 |
| 3 | Competition | 82 |
| 4 | Career | 46 |
| 5 | Course | 36 |
| 6 | Masterclass | 22 |
| 7 | Engageme... | 20 |

**5. Duplicate Opportunity Codes** sql

CopyEdit

SELECT opportunity_code, COUNT(*)

FROM "Opportunity_Raw" GROUP

BY opportunity_code

HAVING COUNT(*) > 1;

**Output:**

| | opportunity_code text | count bigint |
|---|---|---|
| 1 | MDL3K7I | 2 |
| 2 | E207779 | 2 |
| 3 | E443361 | 2 |
| 4 | E8B2TZ7 | 2 |
| 5 | I860340 | 2 |
| 6 | A9Q15OS | 2 |
| 7 | E352968 | 2 |
| 8 | ATGVKWF | 2 |
| 9 | A209733 | 2 |
| 10 | EBOGN8J | 2 |
| 11 | A1K1ITG | 2 |
| 12 | MGIK3RO | 2 |
| 13 | MCTI4XR | 2 |
| 14 | IBLCQ1D | 2 |
| 15 | A2S2WSK | 2 |
| 16 | AUCX0B3 | 2 |
| 17 | A7SZRN9 | 2 |
| 18 | ER044NC | 2 |
| 19 | ICKXVLL | 2 |
| 20 | I25190A | 2 |
| 21 | MAGWCSP | 2 |

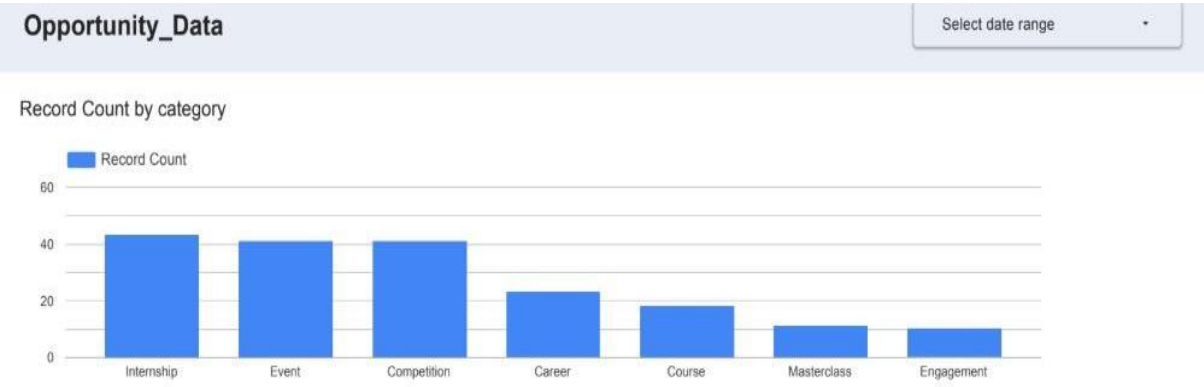
### 6. Missing value after cleaning

```
SELECT
COUNT(*) AS total_rows,
COUNT(*) - COUNT(opportuniy_id) AS missing_opportuniy_id,
COUNT(*) - COUNT(opportuniy_name) AS missing_opportuniy_name,
COUNT(*) - COUNT(category) AS missing_category,
COUNT(*) - COUNT(opportunity_code) AS missing_opportunity_code,
COUNT(*) - COUNT(tracking_questions) AS missing_tracking_questions
FROM "Opportunity_Raw";
```

**Output:**

| | total_rows bigint | missing_opportuniy_id bigint | missing_opportuniy_name bigint | missing_category bigint | missing_opportunity_code bigint | missing_tracking_questions bigint |
|---|---|---|---|---|---|---|
| 1 | 374 | 0 | 0 | 0 | 0 | 0 |

The Opportunity dataset reveals that most listed opportunities fall under internships, followed closely by events and competitions. Learners are primarily engaging with opportunities related to cybersecurity, data visualization, and project management—topics that reflect high demand for technical and career advancement skills. This trend suggests a strong preference among learners for hands-on, skill-based learning experiences.




EDA_REPORT Table

| | opportunity_name | Record Count |
|---|---|---|
| 1. | Cybersecurity: Defensive Hacking | 4 |
| 2. | Data Visualization | 4 |
| 3. | Project Management | 4 |
| 4. | Digital Marketing | 3 |
| 5. | Jump Start: Developing your Emotional Intelligence | 2 |
| 6. | Million Dollar Idea | 2 |
| 7. | AI Forensic Challenge | 2 |

1 - 100 / 169

- **Most opportunities are internships and courses, followed by events and competitions.**
- **Learners are most engaged with topics like cybersecurity, data visualization, and project management.**

**Conclusion:**
The analysis of the Opportunity_Raw dataset revealed that while most key fields are wellpopulated, some columns—especially tracking_questions—contain missing or empty values. A few duplicate opportunity_code entries were also identified. These issues should be addressed to ensure accurate reporting and better data reliability for further analysis or integration.

The **LearnerOpportunity_Raw** dataset captures learner enrollments across various cohorts, along with enrollment details like status, apply_date, and enrollment_id. The purpose of this EDA was to understand the structure of the data, assess data quality, and uncover patterns in learner participation and cohort assignment.

The analysis began with basic data profiling, including checking row count, completeness of key columns, and uniqueness of values. Further queries were used to explore how learners are distributed across cohorts, analyze enrollment statuses, and flag potential duplicates.

### 1. Data Preview

```
SELECT * FROM "LearnerOpportunity_Raw" LIMIT 10;
```

**Displays the first 10 records for an overview of the structure and values.**

| | 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 |

### 2. Missing Values Check

```
SELECT
  COUNT(*) AS total_rows,
  COUNT("enrollment_id") AS non_null_enrollment_id,
  COUNT("learner_id") AS non_null_learner_id,
  COUNT("assigned_cohort") AS non_null_assigned_cohort,
  COUNT("apply_date") AS non_null_apply_date,
  COUNT("status") AS non_null_status
FROM "LearnerOpportunity_Raw";
```

**Identifies completeness of each column by showing non-null counts.**

| | total_rows bigint | non_null_enrollment_id bigint | non_null_learner_id bigint | non_null_assigned_cohort bigint | non_null_apply_date bigint | non_null_status bigint |
|---|---|---|---|---|---|---|
| 1 | 113602 | 113602 | 113602 | 113602 | 113602 | 113602 |

**3. Distinct Count of Key Fields**

```
SELECT
  COUNT(DISTINCT "learner_id") AS unique_learners,
  COUNT(DISTINCT "assigned_cohort") AS unique_cohorts,
  COUNT(DISTINCT "status") AS unique_statuses
FROM "LearnerOpportunity_Raw";
```

**Measures uniqueness in learners, cohorts, and status types.**

| | total_rows bigint | non_null_enrollment_id bigint | non_null_learner_id bigint | non_null_assigned_cohort bigint | non_null_apply_date bigint | non_null_status bigint |
|---|---|---|---|---|---|---|
| 1 | 113602 | 113602 | 113602 | 113602 | 113602 | 113602 |

**4. Status Distribution**

```
SELECT "status", COUNT(*) AS count
FROM "LearnerOpportunity_Raw"
GROUP BY "status"
ORDER BY count DESC;
```

**Helps understand how many learners fall under each enrollment status.**

**Output:**

| | status text | count bigint |
|---|---|---|
| 1 | 1070 | 76109 |
| 2 | 1030 | 12236 |
| 3 | 1055 | 11471 |
| 4 | 1120 | 9048 |
| 5 | 1110 | 1514 |
| 6 | 1080 | 1191 |
| 7 | 1050 | 1003 |
| 8 | 1010 | 659 |
| 9 | NULL | 186 |
| 10 | 1020 | 161 |
| 11 | 1040 | 24 |

**5. Cohort-wise Learner Distribution**

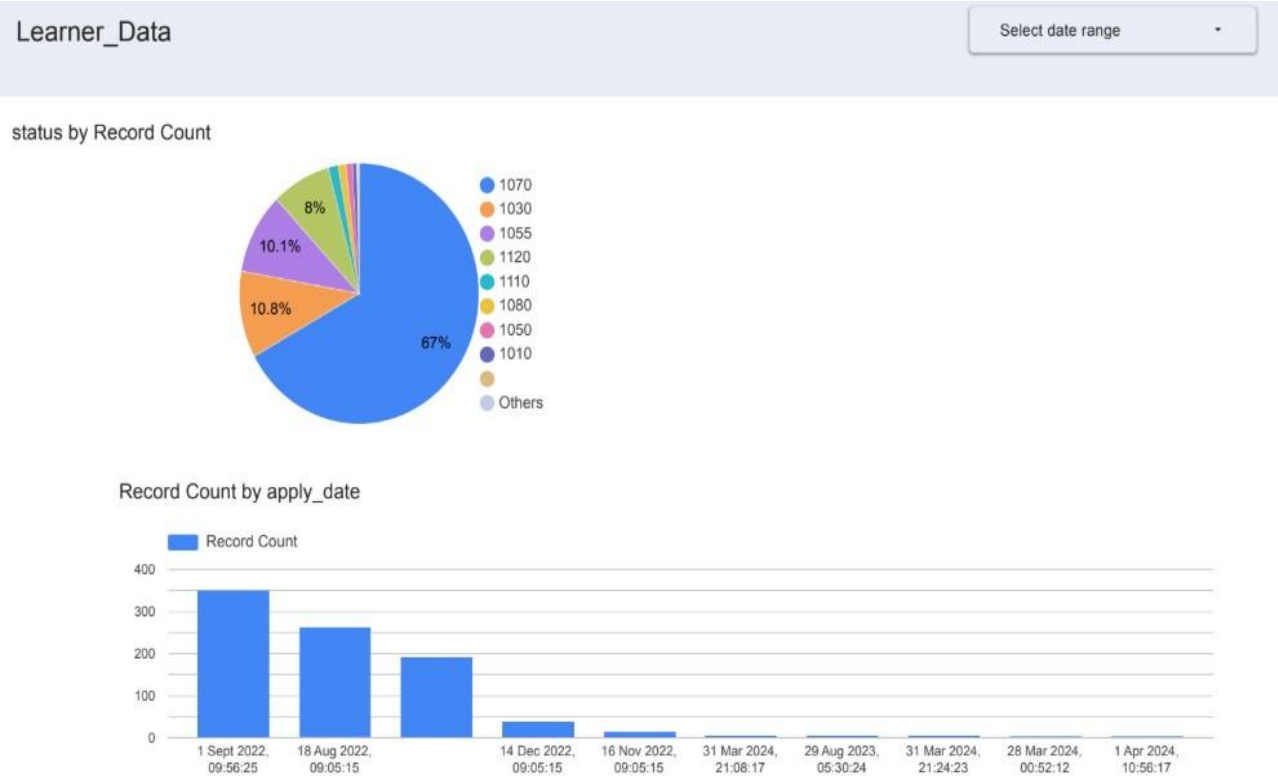
```
SELECT "assigned_cohort", COUNT(*) AS learner_count
FROM "LearnerOpportunity_Raw"
GROUP BY "assigned_cohort"
ORDER BY learner_count DESC;
```

**Reveals which cohorts have the highest number of learners assigned.**

| | assigned_cohort text | learner_count bigint |
|---|---|---|
| 1 | NULL | 13318 |
| 2 | BAM6HBR | 1805 |
| 3 | BSEV9QO | 1733 |
| 4 | BGRQZ2N | 1719 |
| 5 | BP9ZV19 | 1611 |
| 6 | BWAG78I | 1564 |
| 7 | BT4YTCR | 1532 |
| 8 | BEXFE8O | 1525 |
| 9 | B986905 | 1522 |
| 10 | B6MZ4HK | 1502 |
| 11 | BE7X8PZ | 1497 |
| 12 | BPPMQ44 | 1437 |
| 13 | BJZ6HXM | 1309 |
| 14 | BKXA704 | 1112 |
| 15 | BSQOBO9 | 1039 |
| 16 | BIYM5IR | 988 |
| 17 | BN42GTI | 905 |
| 18 | B73QELN | 881 |
| 19 | BG3G48P | 872 |
| 20 | B334245 | 863 |
| 21 | B9FU29Y | 837 |

# C. LearnerOpportunity_Raw

The **LearnerOpportunity** dataset shows that the majority of learners (67%) are categorized under status code 1070, with smaller proportions under 1030 (10.8%) and 1055 (10.1%), indicating a concentration of learners at a particular enrolment stage. Application activity peaked on 1st September 2022 and 18th August 2022, after which there was a significant drop, especially post-December 2022. This suggests that learner engagement was highest during that period, and formatting apply dates to Month-Year could enhance clarity in visualizations.




- **Status 1070 dominates with 67% of total records, followed by 1030 (10.8%) and 1055 (10.1%).**
- **These numeric codes likely represent learner states (e.g., enrolled, applied).**
- **Application activity peaked on 1 Sept 2022 and 18 Aug 2022.**
- **Few applications are recorded after Dec 2022. • Apply dates include timestamps, making the chart dense — suggest formatting to show Month-Year.**

**Conclusion:**

The dataset appears structurally sound with most fields well populated. A significant number of learners are assigned to a few key cohorts, and the status field provides useful insights into application stages. However, a few duplicate entries, where learners appear multiple times in the same cohort, were found and should be reviewed. Overall, the dataset is mostly clean and provides meaningful information about learner engagement across learning opportunities.

# D. Cognito User Data Analysis:

The Cognito_Raw2 dataset contains user profile and account metadata, including demographic attributes such as gender, birthdate, and location (city, state, zip), along with system-generated fields like UserCreateDate and UserLastModifiedDate. This exploratory data analysis aims to assess data quality, uniqueness of user accounts, demographic distribution, and anomalies in account creation or modification timelines. Understanding these aspects is crucial for user segmentation, personalization, and improving platform experience.

**1. Check for Missing Values in Key Fields**

```
SELECT
COUNT(*) AS total_rows,
COUNT(user_id) AS non_null_user_id,
COUNT(email) AS non_null_email,
COUNT(gender) AS non_null_gender,
COUNT("UserCreateDate") AS non_null_created,
COUNT("UserLastModifiedDate") AS non_null_modified,
COUNT(birthdate) AS non_null_birthdate,
COUNT(city) AS non_null_city,
COUNT(zip) AS non_null_zip,
COUNT(state) AS non_null_state
FROM "Cognito_Raw2";
```

**Output:**

| | total_rows bigint | non_null_user_id bigint | non_null_email bigint | non_null_gender bigint | non_null_created bigint | non_null_modified bigint | non_null_birthdate bigint | non_null_city bigint | non_null_zip bigint |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 34111 | 34111 | 34111 | 34111 | 34111 | 34111 | 34111 | 34111 | 341 |

**2. Identify extract the year from birthdate**

**SELECT EXTRACT(YEAR FROM birthdate) AS birth_year, COUNT(*)**

**FROM "Cognito_Raw2"**

**GROUP BY birth_year ORDER BY birth_year;**
**Output:**

| | birth_year numeric | count bigint |
|---|---|---|
| 1 | 1925 | 1 |
| 2 | 1935 | 1 |
| 3 | 1942 | 1 |
| 4 | 1958 | 2 |
| 5 | 1960 | 1 |
| 6 | 1961 | 1 |
| 7 | 1963 | 1 |
| 8 | 1964 | 2 |
| 9 | 1965 | 3 |
| 10 | 1966 | 1 |
| 11 | 1967 | 5 |
| 12 | 1968 | 2 |
| 13 | 1969 | 4 |
| 14 | 1970 | 7 |
| 15 | 1971 | 4 |
| 16 | 1972 | 9 |
| 17 | 1973 | 14 |
| 18 | 1974 | 16 |
| 19 | 1975 | 23 |
| 20 | 1976 | 28 |

**3. Gender Distribution**

SELECT gender, COUNT(*) AS count

FROM "Cognito_Raw2"

GROUP BY gender

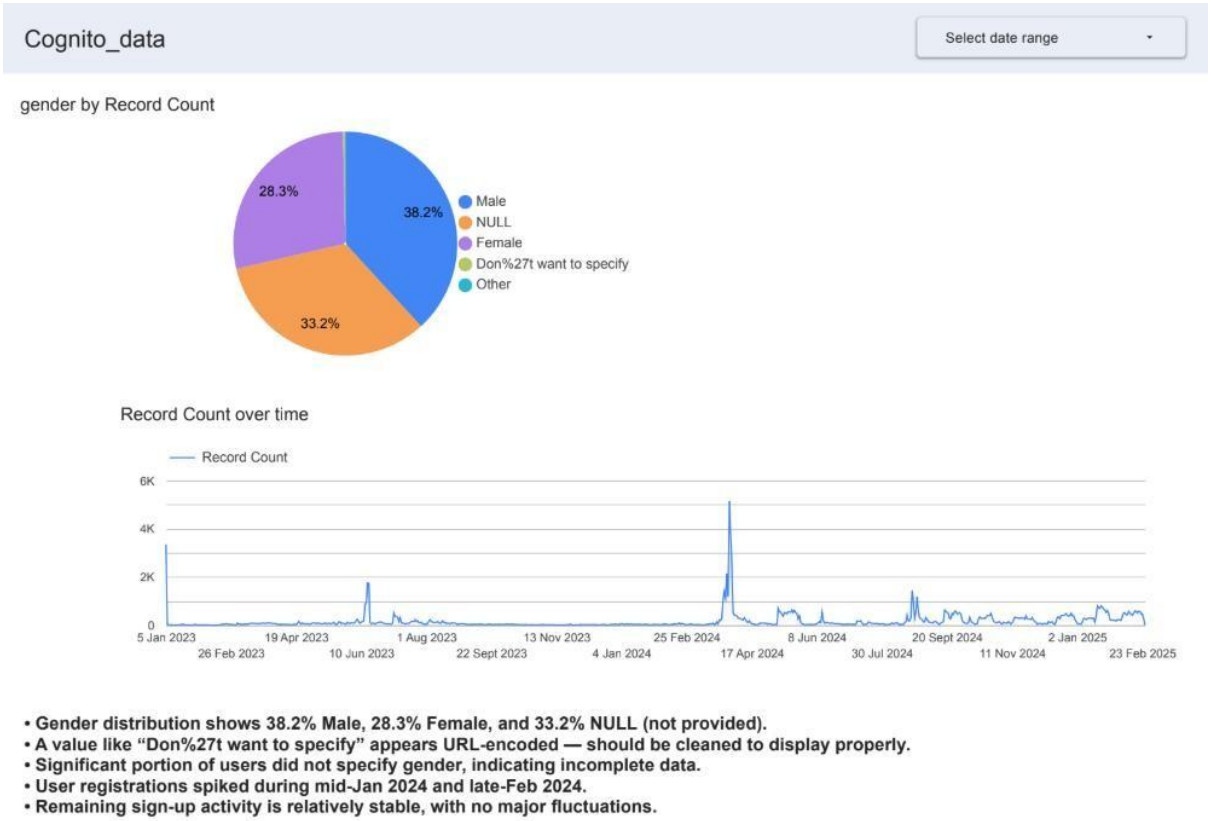
ORDER BY count DESC;

**Output:**

| | gender text | count bigint |
|---|---|---|
| 1 | Male | 19890 |
| 2 | Female | 14046 |
| 3 | Don%27t want to specify | 145 |
| 4 | Other | 30 |

# D. Cognito User Data Analysis:

The Cognito data dashboard reveals a gender distribution where 38.2% of users identify as male, 28.3% as female, and 33.2% have not provided gender information (NULL), indicating a notable gap in demographic completeness. An additional value, “Don%27t want to specify,” appears URL-encoded and should be cleaned for clarity. The time-series chart shows stable registration trends, with noticeable spikes in user sign-ups during mid-January and lateFebruary 2024. Overall, the dataset suggests steady platform engagement, but highlights the need for gender data validation and cleanup to ensure more accurate profiling.




- **Gender distribution shows 38.2% Male, 28.3% Female, and 33.2% NULL (not provided).**
- **A value like “Don%27t want to specify” appears URL-encoded — should be cleaned to display properly.**
- **Significant portion of users did not specify gender, indicating incomplete data.**
- **User registrations spiked during mid-Jan 2024 and late-Feb 2024.**
- **Remaining sign-up activity is relatively stable, with no major fluctuations.**

## Conclusion:

The analysis of the Cognito_Raw2 dataset highlighted a mostly complete set of user profiles, though some fields such as birthdate had missing values. Duplicate email entries were detected, suggesting the need for validation to maintain unique user identities. Gender distribution analysis can aid in demographic profiling, while the detection of records where the UserLastModifiedDate precedes UserCreateDate indicates potential data entry or system syncing errors. These insights can support improved data integrity and user management strategies.

# E. Cohort Raw

The **CohortRaw** table contains cohort-related metadata such as IDs, codes, start/end dates, and size. The primary goal of this analysis is to clean and extract usable numeric cohort IDs, validate date logic, assess missing values, and analyse cohort sizes to ensure consistency and readiness for downstream use.

**1. Check for Missing & Non-Null Values**

```
SELECT
  COUNT(*) AS total_rows,
  COUNT("cohort_id") AS non_null_cohort_id,
  COUNT("cohort_code") AS non_null_cohort_code,
  COUNT("start_date") AS non_null_start,
  COUNT("end_date") AS non_null_end,
  COUNT("size") AS non_null_size
FROM "CohortRaw"; Output:
```

| | total_rows bigint | non_null_cohort_id bigint | non_null_cohort_code bigint | non_null_start bigint | non_null_end bigint | non_null_size bigint |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |

**2. Validate Date Consistency**

```
SELECT COUNT(*) AS column_count
FROM information_schema.columns
WHERE table_name = 'CohortRaw'; Output:
```

| | column_count bigint |
|---|---|
| 1 | 6 |

**3. Clean and Extract Numeric Cohort ID**

```
UPDATE "CohortRaw"
SET clean_cohort_id = CAST(SUBSTRING("cohort_id" FROM '[0-9]+') AS INTEGER);
```

**Output:** update 0

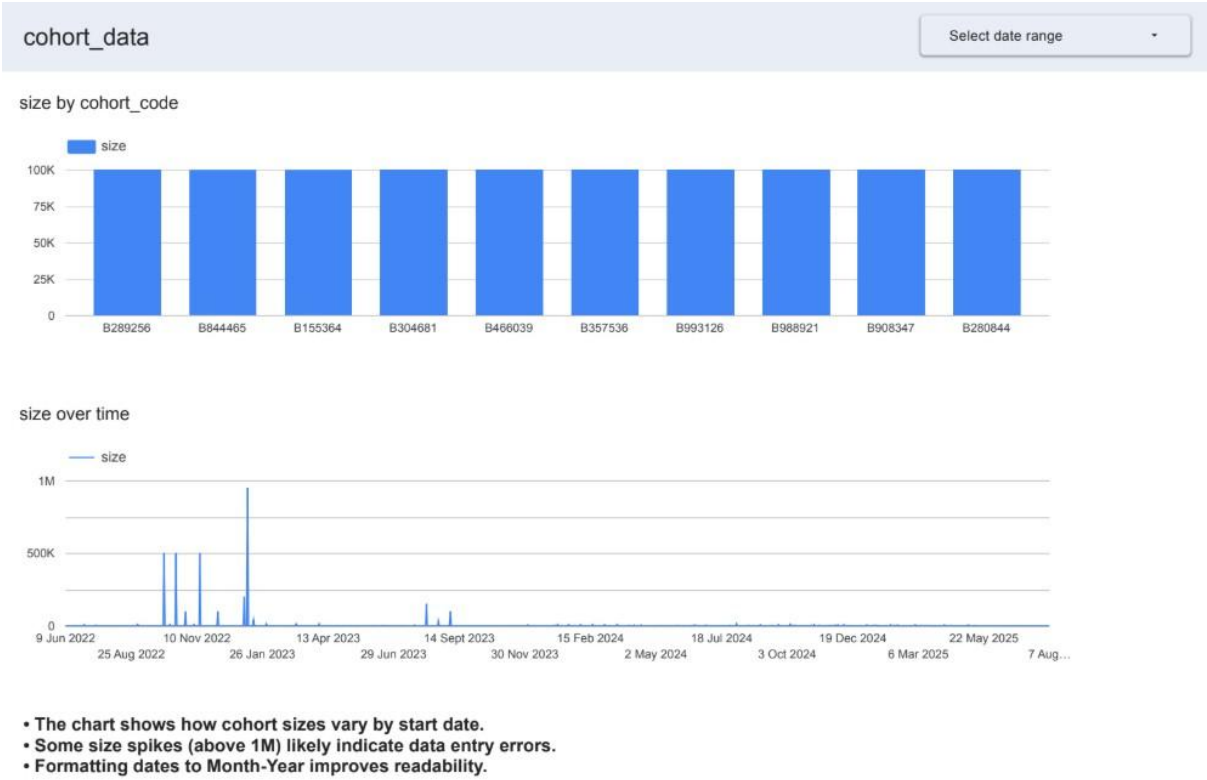
**4. Cohort Size Stats**

SELECT MAX("size"), MIN("size"), AVG("size"), STDDEV("size")

FROM "CohortRaw"; Output:

| | max integer | min integer | avg numeric | stddev numeric |
|---|---|---|---|---|
| 1 | [null] | [null] | [null] | [null] |

The cohort_data dashboard presents cohort sizes by unique codes and over time. Cohort sizes appear relatively consistent across codes, hovering around 100K, suggesting uniform batch grouping. However, the time-based trend reveals unusual spikes exceeding 1 million, which are likely data entry errors or system glitches. These anomalies distort overall trends and should be corrected for accurate analysis. Additionally, improving date formatting to Month-Year would enhance readability.




- **The chart shows how cohort sizes vary by start date.**
- **Some size spikes (above 1M) likely indicate data entry errors.**
- **Formatting dates to Month-Year improves readability.**

**Conclusion**

The CohortRaw dataset required the extraction of clean numeric cohort IDs from string-based cohort_id fields using regex and casting. Minor issues were found in date consistency, such as end dates occurring before start dates, which should be corrected. Several fields had missing values, particularly in dates and size. Statistical analysis of the size field helped identify the range and variation across cohorts, providing useful insights for capacity planning and resource allocation.

# F. Marketing Campaign Data (2023–2024)

The "Marketing Campaign Data All Accounts (2023–2024)" dataset contains key performance metrics across various advertising campaigns, including campaign names, delivery status, reach, clicks, cost per result, and amount spent in AED. This data is essential for evaluating the effectiveness of digital marketing efforts across multiple ad accounts. The goal of this analysis is to extract meaningful insights such as total ad spend, campaign performance by clicks, and cost efficiency, helping to guide future marketing strategies.

**1. Total Ad Spend**

SELECT SUM("Amount spent (AED)") AS total_spent

FROM "Marketing Campaign Data All Accounts (2023-2024)"; Calculates

**the total amount spent across all campaigns.**

**Output:**

| | total_spent bigint |
|---|---|
| 1 | [null] |

**2. Top 5 Campaigns by Outbound Clicks**

SELECT "Campaign name", SUM("Outbound clicks") AS total_clicks

FROM "Marketing Campaign Data All Accounts (2023-2024)"

GROUP BY "Campaign name"

ORDER BY total_clicks DESC

LIMIT 5;

**Identifies the most engaging campaigns based on user click-through behavior.**

The analysis of the Marketing_Data dashboard reveals that the "March Brand Awareness" and "iBrand Awareness" campaigns were the most successful, each generating over 100 million results, making them standout performers in terms of reach. However, data quality issues are evident, one campaign entry lacks a name (null), indicating potential gaps in data input or tracking. The EDA_REPORT Table provides crucial cost-effectiveness metrics such as Cost Per Click (CPC) and engagement indicators like outbound and landing interactions. For instance, the "CPR - March Ads" campaign shows a low CPC of 0.52 AED with significant outbound clicks and landing pages, suggesting efficient performance. However, other campaigns such as the one with a CPC of 1.43 AED may reflect higher costs with relatively moderate engagement. To enhance insights further, it's recommended to introduce a time filter (e.g., Reporting Starts) for temporal trend analysis and better optimization of future campaigns.


EDA_REPORT Table

| | Campaign name | Amount spent (AED) | CPC (... | Outbou... | Landing ... |
|---|---|---|---|---|---|
| 1. | March Brand Awareness | 25,531.47 | 0.77 | 32,289 | 5,584 |
| 2. | CPR - March Ads: Website Leads Prospecting \| 18 to ... | 19,756.78 | 0.52 | 38,284 | 16,060 |
| 3. | null | 19,330.51 | 1.43 | 17,494 | 12,214 |
| 4. | Dec. INTERNSHIP: Virtual Internship | 14,935.46 | 0.88 | 16,961 | 12,387 |
| 5. | JAN. INTERNSHIP: Virtual Internship | 14,887.13 | 0.54 | 27,483 | 21,140 |
| 6. | Career Essentials - March Ads: Website Leads Prospe... | 14,337.48 | 1.24 | 11,508 | 4,826 |
| 7. | Jan 2025 \| Masterclass \| SBD \| Dust Challenge | 13,252.91 | 0.94 | 14,090 | 6,700 |

1 - 100 / 138

- **March Brand Awareness and iBrand Awareness delivered the highest results, exceeding 100M+.**
- **Some values appear inflated or incomplete — one row has a missing campaign name.**
- **The table shows CPC and engagement metrics, useful for analyzing cost-effectiveness.**
- **Suggest adding a time filter (Reporting Starts) for trend insights.**

## Conclusion:

The marketing dataset provides clear insights into campaign effectiveness over the 2023–2024 period. The total ad spend highlights the overall budget investment, while outbound clicks and cost-per-result metrics reveal which campaigns delivered the best engagement and ROI. Analysing these patterns helps marketing teams make data-driven decisions, optimize future campaigns, and allocate resources to high-performing strategies.

# Final Conclusion of the Report

- After analysing all dataset, including learner profiles, opportunities, cohort assignments, engagement timelines, and marketing campaign performance, it is evident that structured data cleaning and exploratory analysis are crucial to unlocking actionable insights. Key patterns emerged: learner activity was concentrated in mid-2022, with certain statuses and cohorts dominating; opportunity engagement revealed a preference for technical and career-focused programs; and the Cognito user data highlighted demographic gaps and areas needing validation. The marketing campaign data showed strong campaign performance variation, helping identify the most cost-effective and engaging strategies.

- This comprehensive EDA enabled detection of data quality issues such as missing values, duplicate records, inconsistent formatting, and timeline anomalies. Through standardization, deduplication, and derived fields (e.g., clean cohort IDs), the data has been prepared for further modeling, reporting, or business intelligence use. These findings lay a solid foundation for enhancing learner targeting, refining content delivery, and improving campaign performance going forward.