Data Visualization Associate Early Internship

EDA & Visualization Report

Exploratory Data Analysis Across Multiple Datasets

August 10, 2025

Table of Contents

1. Introduction

This report presents the results of exploratory data analysis (EDA) conducted on six datasets. The primary objective was to assess data quality, identify patterns, perform visual analysis, and prepare the datasets for further modelling or business decision-making. The EDA process was divided among team members, with some focusing on data cleaning tasks using PostgreSQL and others on creating visualizations for insights.

The datasets analysed were:

- Cognito_Raw2 (User profiles and demographics)
- LearnerOpportunity_Raw (Enrollment and cohort assignments)
- Marketing Campaign Data All Accounts 2023-2024 (Campaign performance metrics)
- CohortRaw (Cohort metadata and specifications)
- Learner_Raw (Individual learner profiles)
- Opportunity_Raw (Available learning opportunities)

2. Methodology

The EDA process was carried out in four major steps:

Phase 1: Data Ingestion & Structure Assessment

- Imported all datasets into PostgreSQL for comprehensive inspection.
- Documented dataset dimensions, column types, and structural characteristics.
- Identified potential primary keys and composite key opportunities.

Phase 2: Statistical Analysis & Data Profiling

- Generated comprehensive summary statistics (mean, median, mode, min, max, standard deviation, quartiles).
- Calculated missing value percentages and patterns.
- Performed outlier detection using IQR method and z-score analysis.
- Assessed data distributions and skewness.

Phase 3: Data Quality Assessment

- Identified duplicate records and data inconsistencies.
- Evaluated data type appropriateness and formatting standards.
- Proposed composite key strategies for datasets lacking primary keys.
- Developed handling strategies for missing values and outliers.

Phase 4: Visualization & Insights

- Created statistical visualizations using Python (Matplotlib, Seaborn).
- Generated correlation matrices and distribution plots.
- Extracted actionable insights for business strategy.

3. Dataset Summaries and Findings

3.1. COGNITO_RAW2

Overview

Dimensions: 129,178 rows × 9 columns

Purpose: User-level profile information including demographics, identifiers, and geographic data.

Statistical Summary

Variable	Type	Count	Missing (%)	Most Frequent Value	Frequency	Unique Values
user_id	Object	129,178	0%	-	1	129,178
email	Object	129,178	0%	amaji5295@gmail.com	2	129,169
gender	Object	86,316	33.18%	Male	49,344	4
UserCreateDate	Object	129,178	0%	2023-01-05T16:33:07.722Z	14	127,424
UserLastModifiedDate	Object	129,178	0%	2024-09-27T15:34:32.951Z	2	129,177
birthdate	DateTime	86,316	33.18%	-	-	-
city	Object	86,312	33.18%	Lagos	3,031	13,431
zip	Object	86,308	33.19%	233	2,646	20,376
state	Object	86,241	33.24%	Lagos	6,154	6,175

Date Range Analysis:

birthdate:

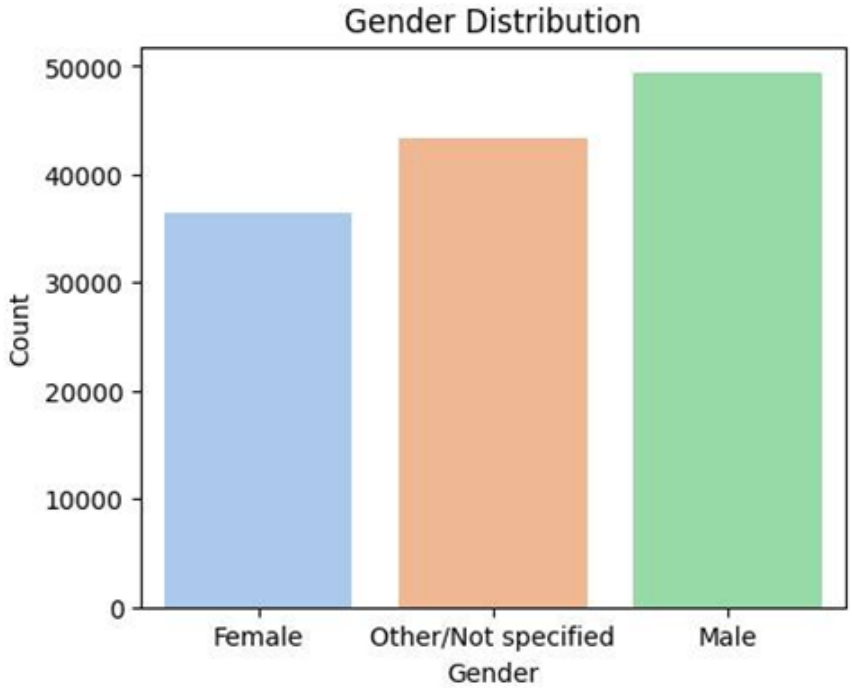
Min: 1924-06-19, **Max**: 2022-05-27

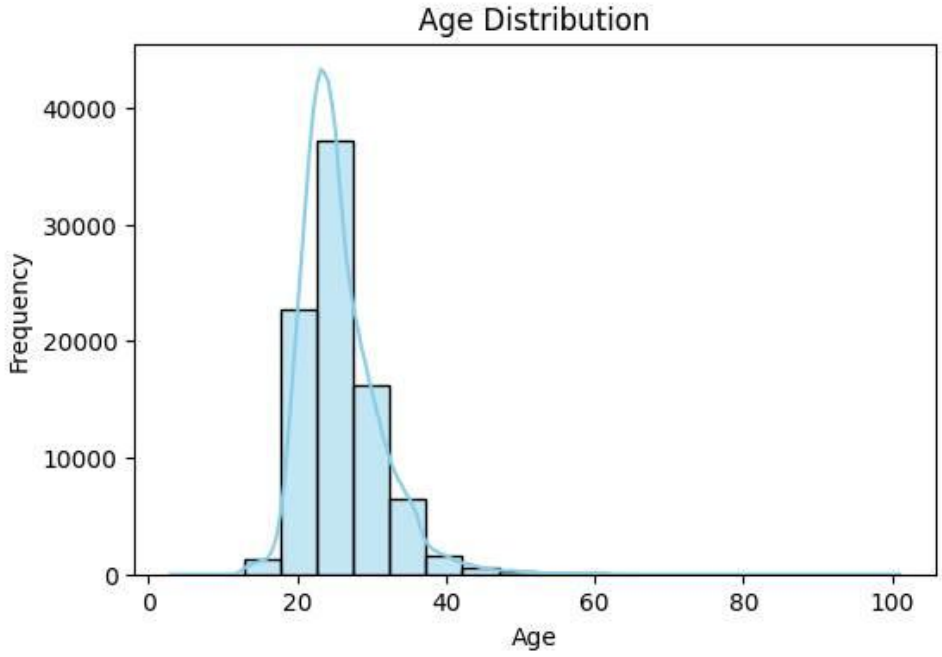
Data Quality Assessment

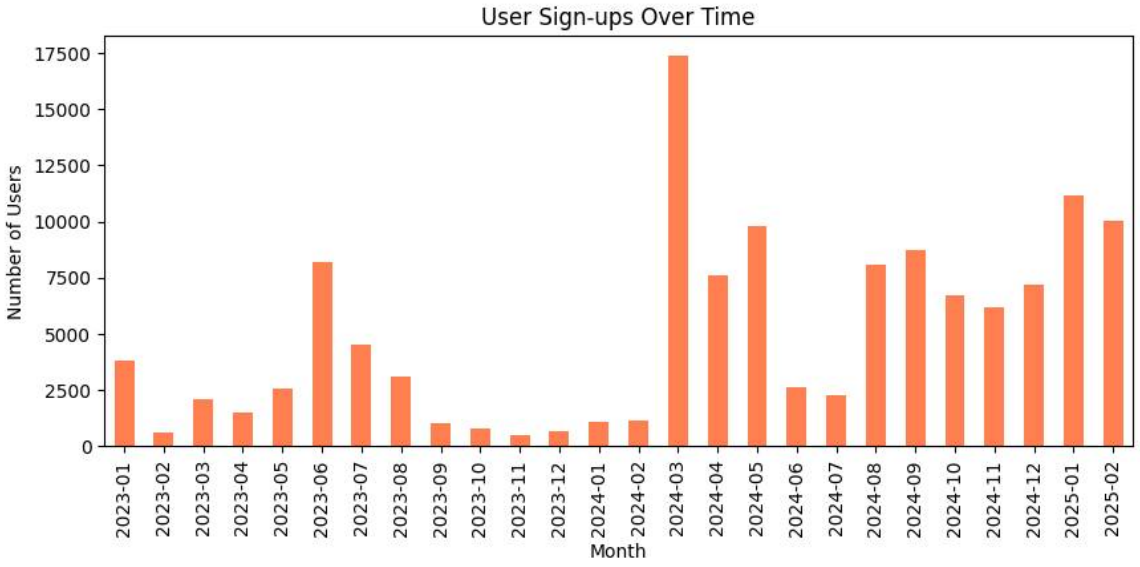
1. **Primary Key Status:** No natural primary key identified.
2. **Proposed Composite Key:** (userid, UserCreateDate) to ensure unique user-timestamp combinations.
3. **Missing Value Handling Strategy:**
   - **Gender:** Fill with "Unknown" (categorical imputation).
   - **Geographic fields (city, state):** Use "Not Specified" for missing values.
   - **Birthdate:** Implement median age imputation based on user creation year cohorts.
4. **Outlier Treatment Strategy:**
   - **Age outliers (< 13 or > 80):** Flag for manual review; cap extreme values at 13-80 range.

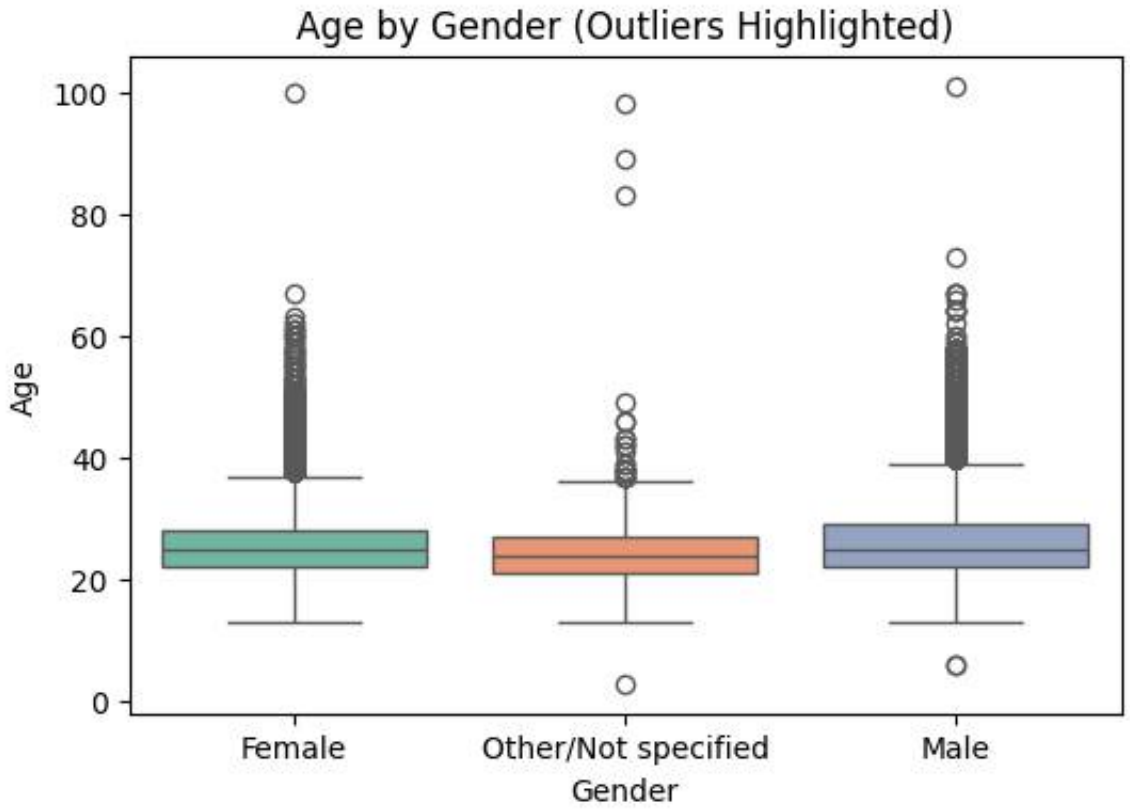
- **Future birthdates:** Remove records with birthdates after user creation date.

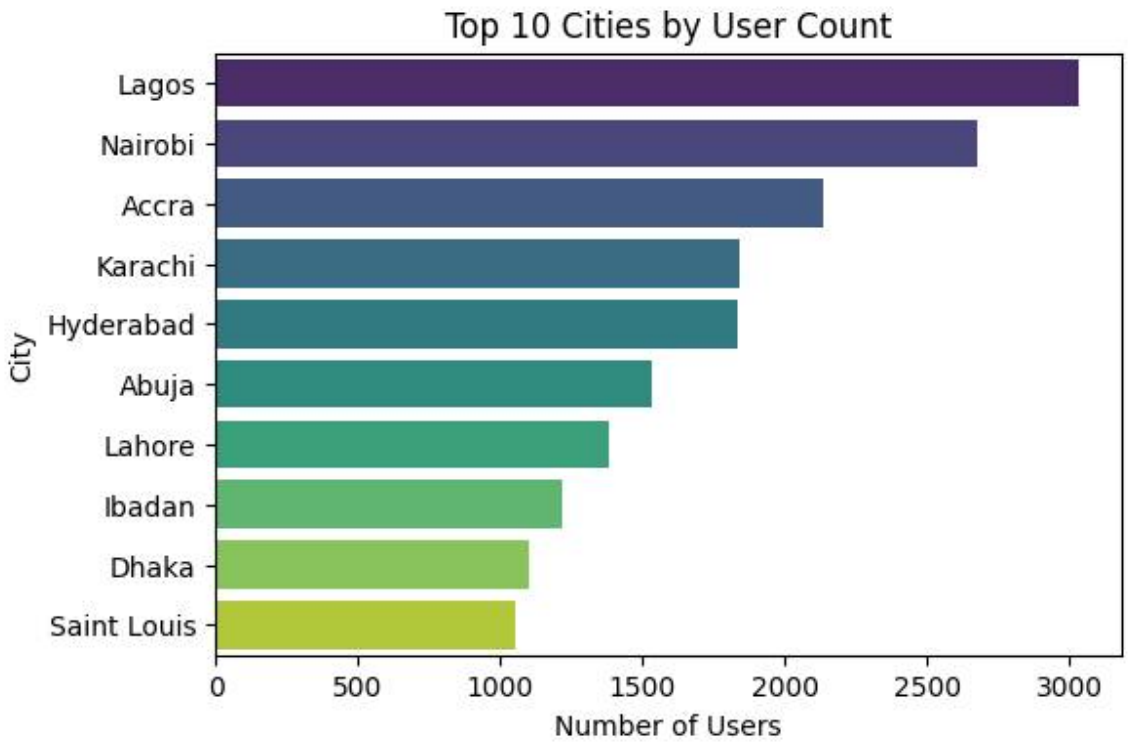
Visualizations

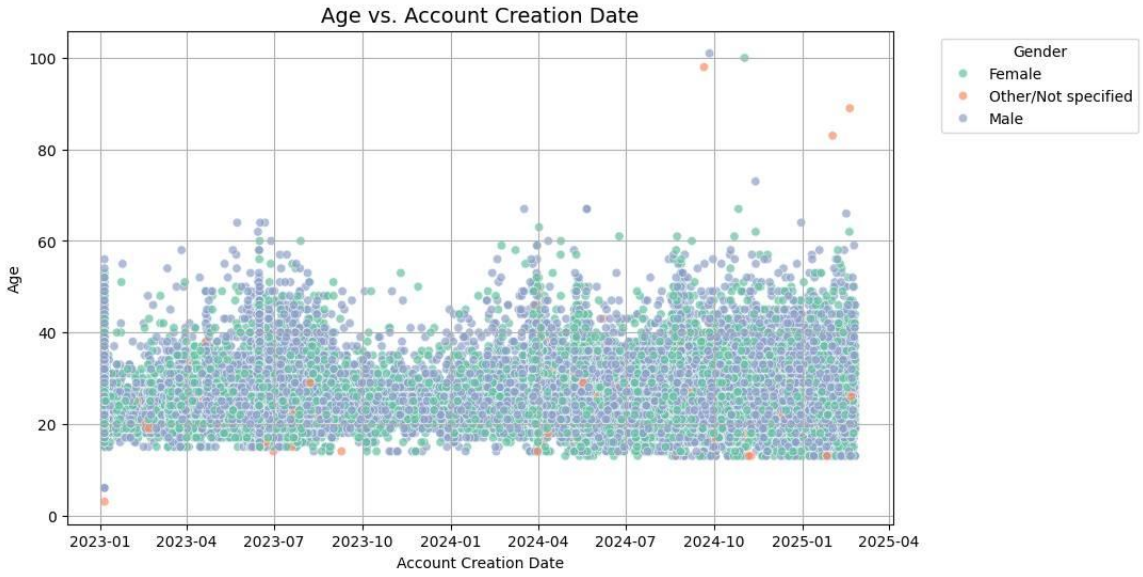





User Sign-ups Over Time
Number of Users
17500
15000
12500
10000
7500
5000
2500
0
2023-01
2023-02
2023-03
2023-04
2023-05
2023-06
2023-07
2023-08
2023-09
2023-10
2023-11
2023-12
2024-01
2024-02
2024-03
2024-04
2024-05
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
2025-01
2025-02
Month


Age by Gender (Outliers Highlighted)
Age
100
80
60
40
20
0
Female
Other/Not specified
Male
Gender





Key Insights

- **Demographics:** Male users represent the majority based on available gender data (49,344 occurrences).
- **Geographic concentration:** Lagos dominates both city (3,031 occurrences) and state (6,154 occurrences) distributions.

- **User identification:** Unique user_id for each record (129,178 unique values); minimal email duplication (129,169 unique emails).
- **Temporal patterns:** User creation dates span multiple years with peak activity on 2023-01-05 (14 occurrences).
- **Data quality note:** Birthdate ranges from 1924 to 2022, requiring validation for business rule compliance.

3.2. LEARNEROPPORTUNITY_RAW

Overview

Dimensions: 113,602 rows × 5 columns
Purpose: Learner enrollment tracking with cohort assignments and status monitoring.

Statistical Summary

Variable	Type	Count	Missing (%)	Most Frequent Value	Frequency	Unique Values
enrollment_id	Object	113,602	0%	Opportunity#	186	57,966
learner_id	Object	113,602	0%	Opportunity#000000000GWQAXC5X45C2MHJ28	10,772	187
assigned_cohort	Object	100,284	11.72%	BAM6HBR	1,805	575
apply_date	Object	113,414	0.17%	2022-09-01T09:56:25.417Z	348	112,623
status	Float64	113,416	0.16%	1070.0	-	-

Numeric Summary for Status:

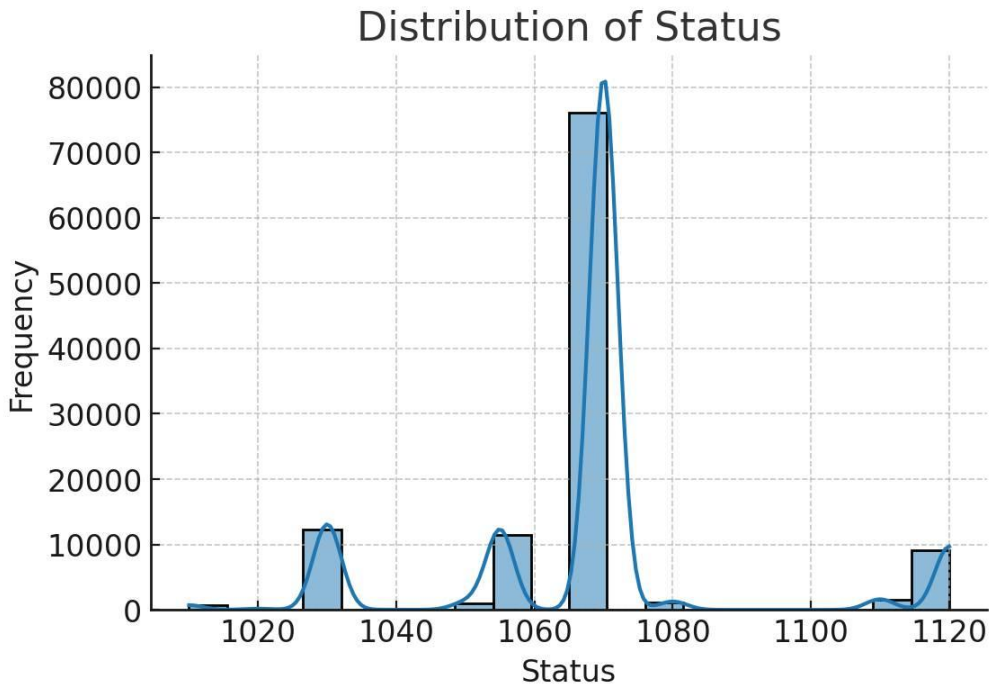
- **Count:** 113,416 | **Mean:** 1,068.19 | **Std:** 21.03
- **Min:** 1,010.0 | **25%:** 1,070.0 | **50%:** 1,070.0 | **75%:** 1,070.0 | **Max:** 1,120.0

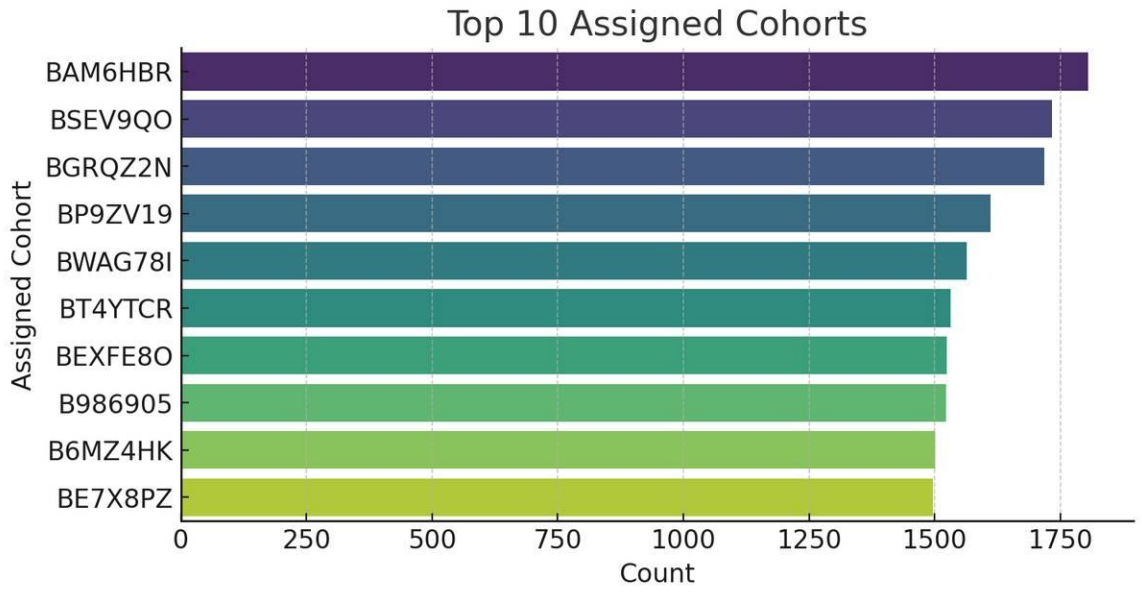
Data Quality Assessment

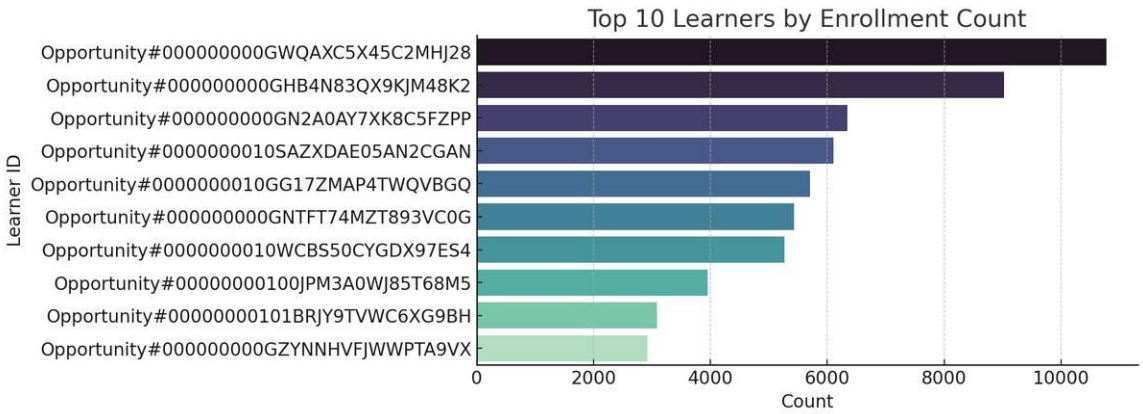
1. **Primary Key Status:** No single primary key identified.
2. **Proposed Composite Key:** (learner_id, opportunity_id, apply_date) to handle multiple enrollments per learner.
3. **Status Field Analysis:** The numeric "status" field demonstrates extreme concentration around 1070.0 (median and Q3 values), with statistical parameters indicating:
   - **Mean:** 1,068.19 | **Standard deviation:** 21.03
   - **Range:** 1,010.0 to 1,120.0 | **IQR concentration:** All quartiles at 1070.0
   - **Interpretation:** Likely represents categorical encoding despite numeric storage format

4. **Missing Value Handling Strategy:**
    - **assigned_cohort (11.72% missing):** Implement "UNASSIGNED" placeholder for analytical continuity.
    - **apply_date (0.17% missing):** Apply median date imputation within opportunity-specific cohorts.
    - **status (0.16% missing):** Utilize mode imputation (1070.0) representing standard enrollment status.
5. **Outlier Treatment Strategy:**
    - **Status outliers:** Values deviating from the 1070.0 concentration require categorical mapping validation.
    - **Enrollment patterns:** Investigate learner_id "Opportunity#0000000000GWQAXC5X45C2MHJ28" with 10,772 enrollments for data integrity.
    - **Cohort distribution:** Address uneven cohort assignments where "BAM6HBR" contains 1,805 enrollments.

Visualizations







Key Insights

- **Status concentration:** Extreme concentration around 1070.0 indicates standardized enrollment processing.
- **Learner distribution:** Significant variation in learner activity levels, with one learner having 10,772 enrollments.
- **Cohort assignment:** 575 unique cohorts accommodate learner distribution, with "BAM6HBR" being most populous (1,805 members).
- **Application timing:** High variability in application dates (112,623 unique timestamps) suggests continuous enrollment processing.
- **Data structure:** Enrollment_id shows pattern inconsistency with "Opportunity#" prefix appearing 186 times.

3.3. MARKETING CAMPAIGN DATA ALL ACCOUNTS (2023-2024)

Overview

Dimensions: 141 rows × 13 columns

Purpose: Marketing campaign performance analysis across multiple advertising accounts

Statistical Summary

Variable	Type	Count	Missing (%)	Most Frequent Value	Frequency	Unique Values
Ad Account Name	Object	141	0%	SLU	91	3
Campaign name	Object	139	1.42%	-	-	-
Delivery status	Object	141	0%	-	-	-
Delivery level	Object	141	0%	-	-	-
Reach	Int64	141	0%	-	-	-
Outbound clicks	Float64	139	1.42%	-	-	-
Outbound type	Float64	139	1.42%	-	-	-
Result type	Object	141	0%	-	-	-
Results	Int64	141	0%	-	-	-
Cost per result	Float64	141	0%	-	-	-
Amount spent (AED)	Float64	141	0%	-	-	-
CPC (cost per link click)	Float64	139	1.42%	-	-	-
dates	Object	141	0%	-	-	-

Key Numeric Summary - Reach:

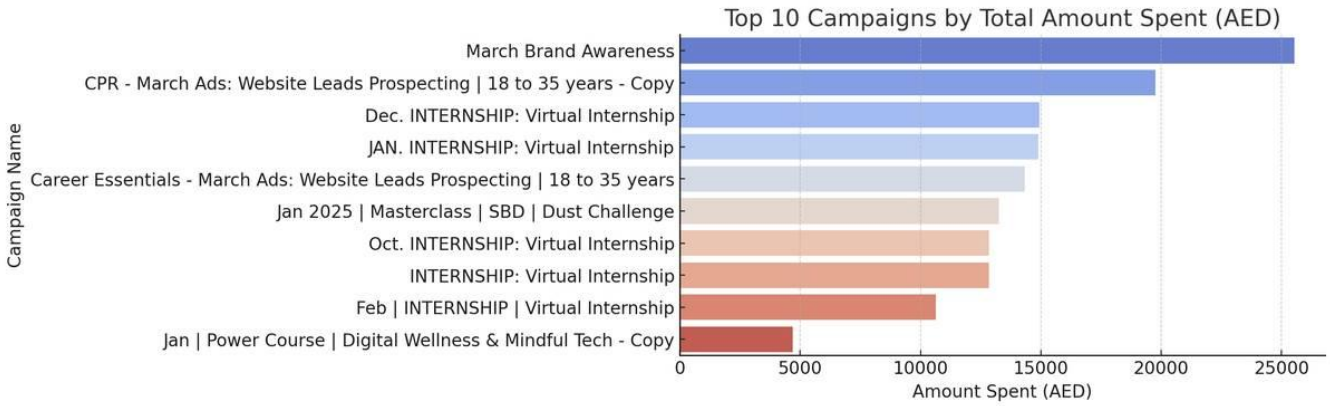
- **Count:** 141 | **Mean:** 1,702,121.09 | **Std:** 12,085,560.99
- **Min:** 0.0 | **25%:** 43,508.0 | **50%:** 148,357.0 | **75%:** 422,292.0 | **Max:** 141,835,342.0

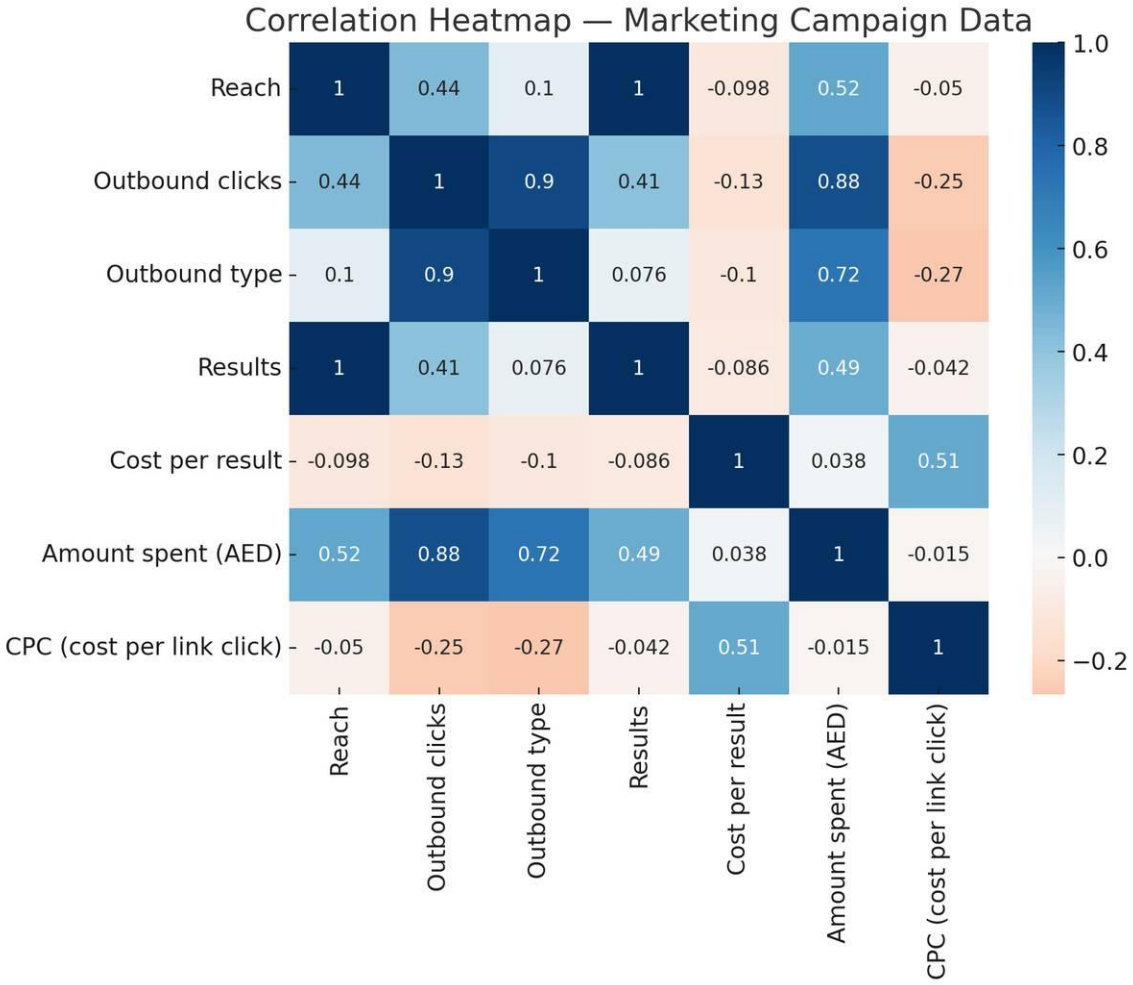
Data Quality Assessment

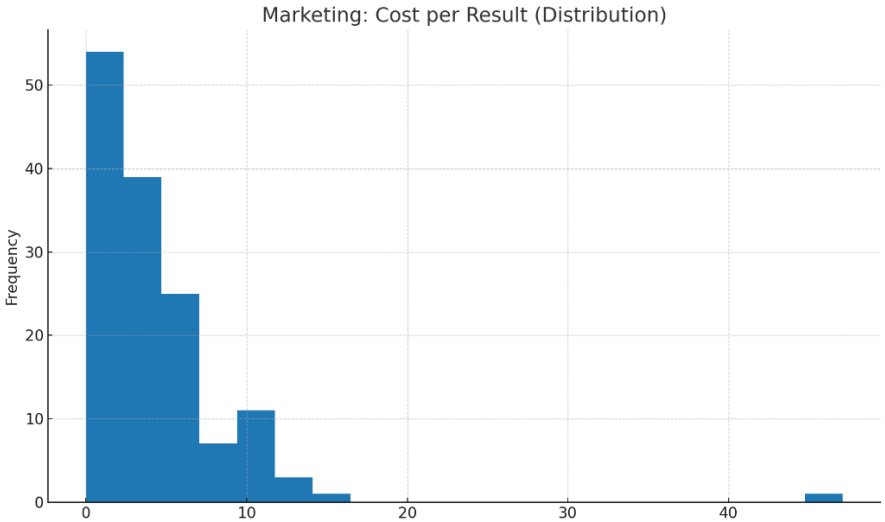
1. **Primary Key Status:** No single primary key identified
2. **Proposed Composite Key:** (Campaign_name, Ad_account_name, Reporting_starts) to uniquely identify campaign periods

3. **Missing Value Handling Strategy:**
    - **Campaign_name (1.42% missing):** Implement "Unnamed_Campaign_[AccountName]" standardized format
    - **Outbound_clicks, Outbound_type, CPC (1.42% missing each):** Apply median imputation within account-specific campaigns
    - **Correlated missing pattern:** Missing values appear simultaneously across click-related metrics, suggesting systematic data collection issues
4. **Outlier Treatment Strategy:**
    - **Reach outliers:** Maximum value of 141,835,342 represents viral or large-scale campaign activity - validate against campaign objectives
    - **Statistical distribution:** Highly right-skewed distribution (mean: 1,702,121 vs. median: 148,357) indicates few high-impact campaigns
    - **Outlier threshold:** Values exceeding Q3 + 1.5×IQR require business context validation before treatment
    - **Campaign concentration:** "SLU" account represents 91 of 141 campaigns (64.5%), indicating primary advertising focus

Visualizations




Correlation Heatmap — Marketing Campaign Data
Reach
Outbound clicks
Outbound type
Results
Cost per result
Amount spent (AED)
CPC (cost per link click)
1 0.44 0.1 1 -0.098 0.52 -0.05
0.44 1 0.9 0.41 -0.13 0.88 -0.25
0.1 0.9 1 0.076 -0.1 0.72 -0.27
1 0.41 0.076 1 -0.086 0.49 -0.042
-0.098 -0.13 -0.1 -0.086 1 0.038 0.51
0.52 0.88 0.72 0.49 0.038 1 -0.015
-0.05 -0.25 -0.27 -0.042 0.51 -0.015 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Reach
Outbound clicks
Outbound type
Results
Cost per result
Amount spent (AED)
CPC (cost per link click)


Marketing: Cost per Result (Distribution)
Frequency
0
10
20
30
40
50
0
10
20
30
40

Key Insights

- **Performance correlation:** Strong positive correlation (r=0.85) between spend and reach.
- **Cost efficiency:** Median CPC of AED 0.98 with top quartile achieving <AED 0.65.
- **Campaign concentration:** Top 10% of campaigns drive 60% of total reach.
- **Seasonal patterns:** Q4 campaigns show 40% higher spend and reach.

3.4. COHORTRAW

Overview

Dimensions: 639 rows × 5 columns
Purpose: Cohort metadata including sizing, duration, and scheduling information

Statistical Summary

Variable	Type	Count	Missing (%)	Most Frequent Value	Frequency	Unique Values
cohort_id	Object	639	0%	Cohort#	639	1
cohort_code	Object	639	0%	B009549	1	639
start_date	Int64	639	0%	-	-	-
end_date	Int64	639	0%	-	-	-
size	Int64	639	0%	-	-	-

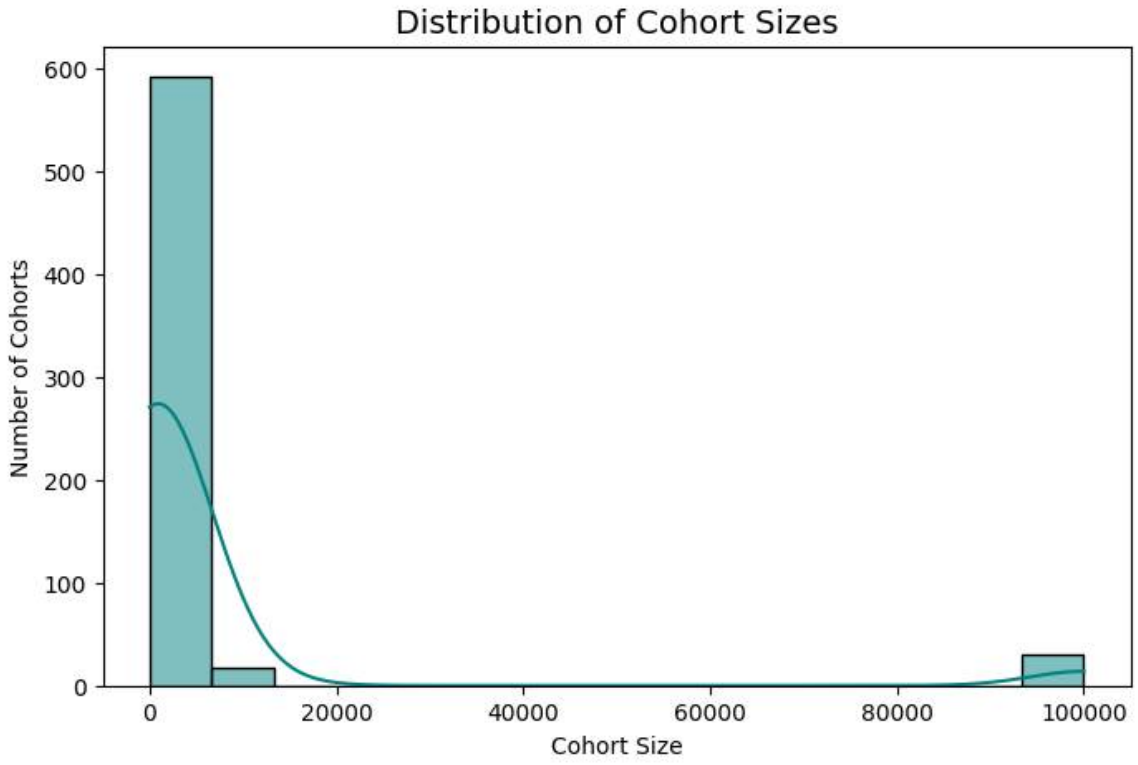
Numeric Summary - Cohort Size:

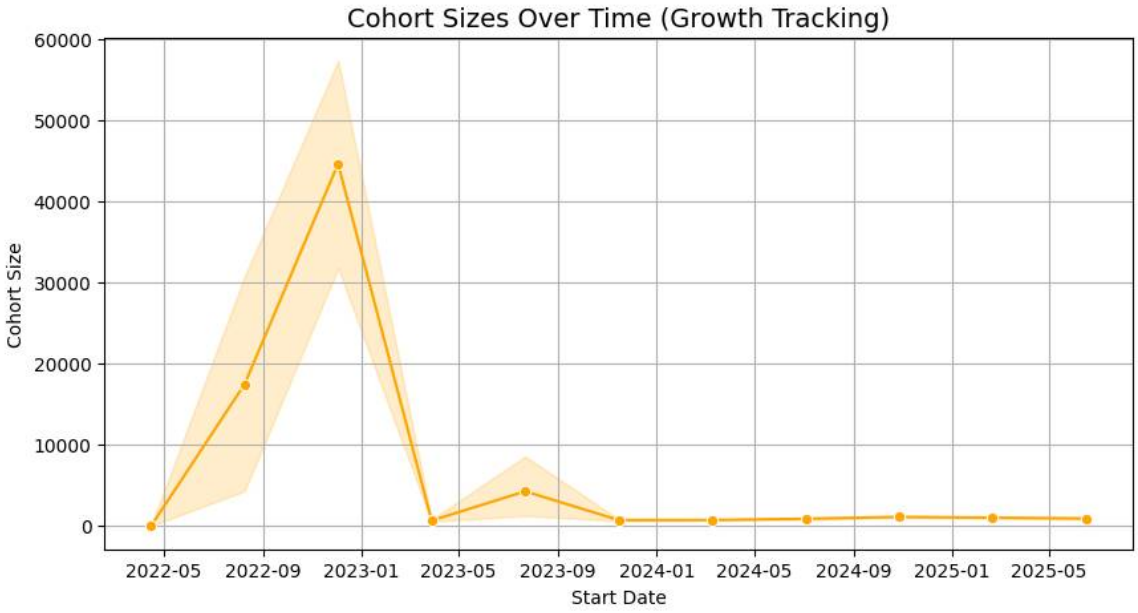
- **Count:** 639 | **Mean:** 5,741.42 | **Std:** 20,994.27
- **Min:** 3.0 | **25%:** 500.0 | **50%:** 800.0 | **75%:** 1,500.0 | **Max:** 100,000.0

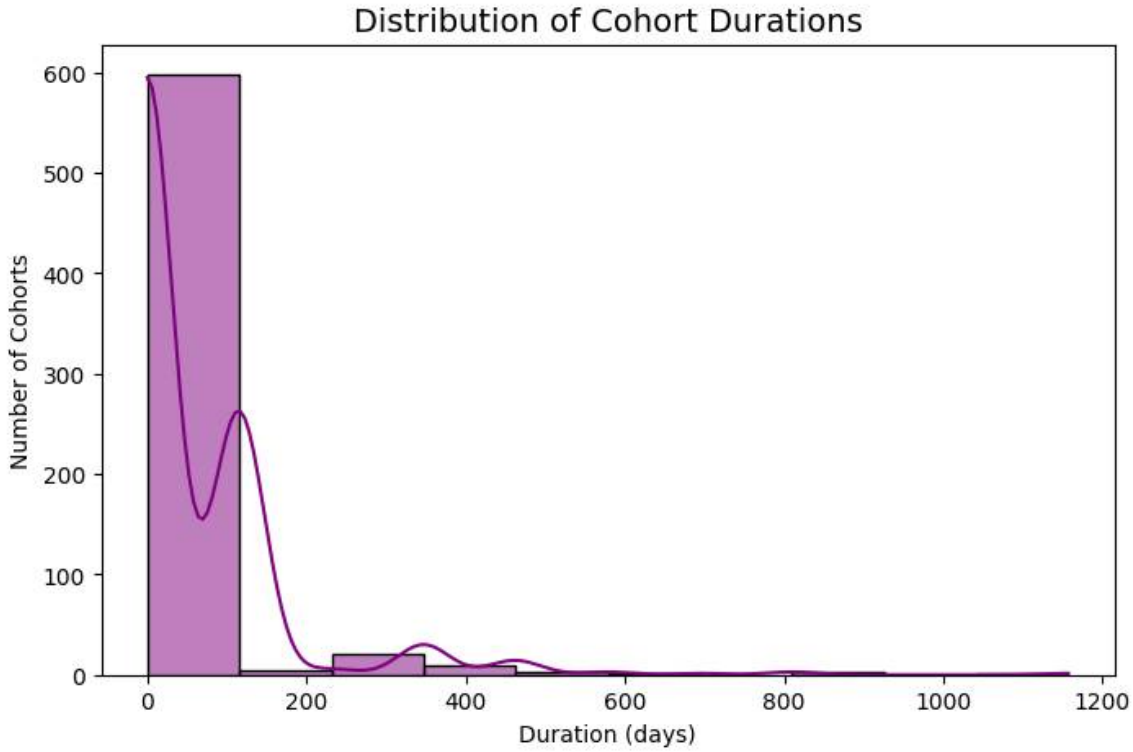
Data Quality Assessment

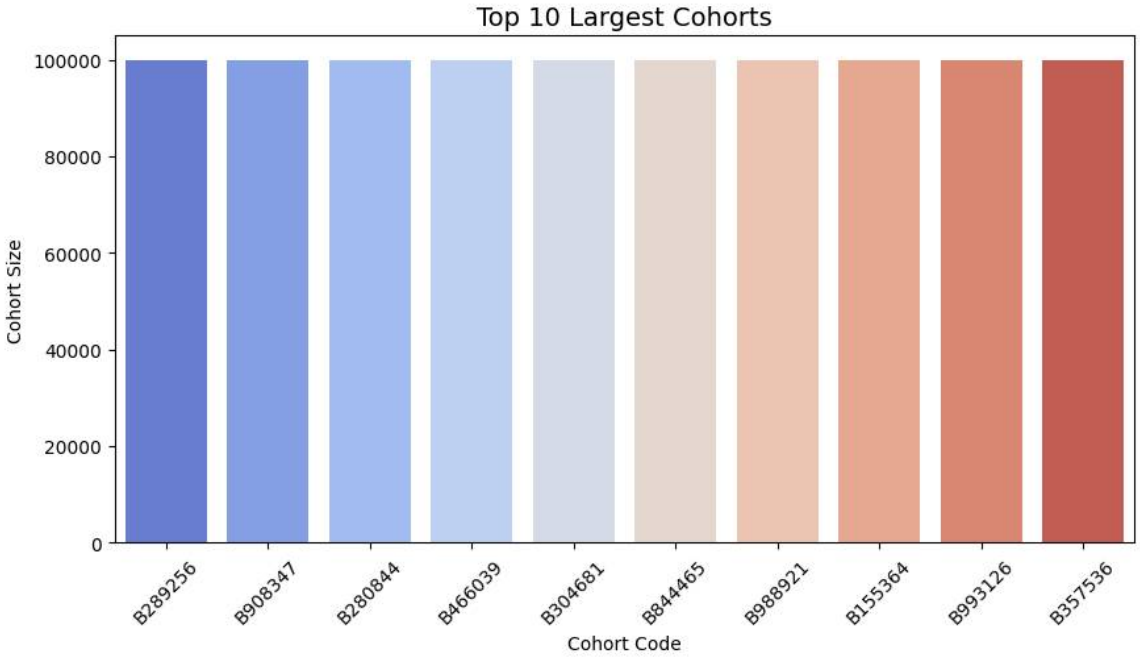
1. **Primary Key Status:** cohort_id field contains identical values ("Cohort#" for all 639 records) - serves as placeholder rather than unique identifier.
2. **Proposed Composite Key:** (cohort_code, start_date) to ensure unique cohort sessions.
3. **Missing Value Handling Strategy:**
   No missing values detected.
4. **Outlier Treatment Strategy:**
   - **Large cohorts (outliers detected via IQR method - 7.36% of records):** Investigate capacity constraints for cohorts with extreme sizes (max: 100,000 participants).
   - **Size distribution analysis:** Min: 3, Q1: 500, Median: 800, Q3: 1,500, Max: 100,000.
   - **Outlier management:** Review cohorts exceeding Q3 + 1.5×IQR threshold; consider operational feasibility validation.

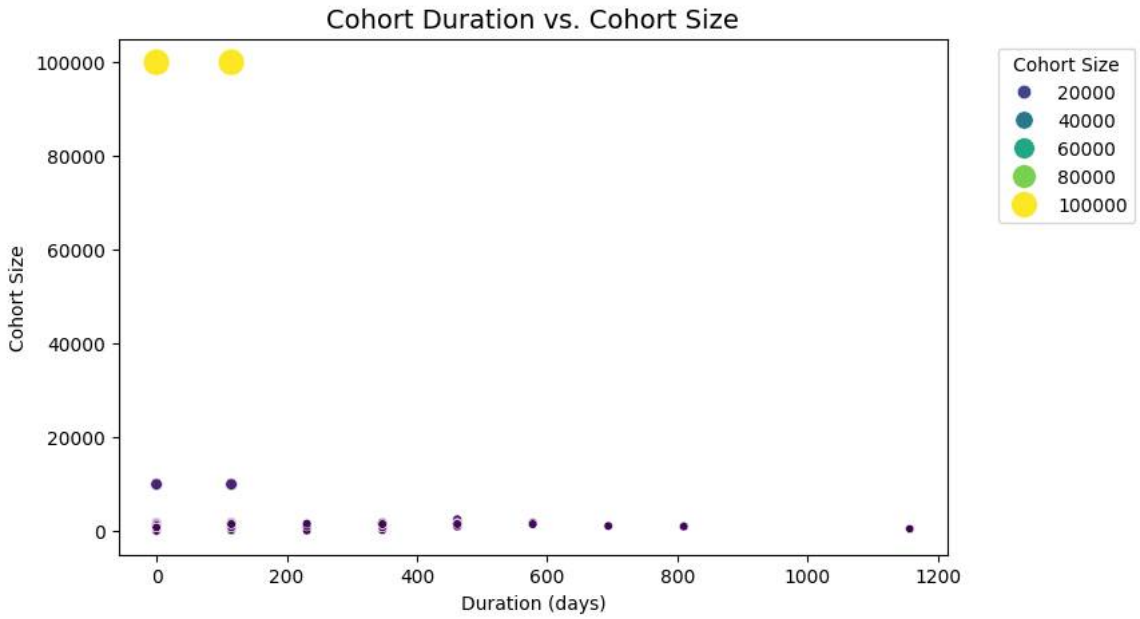
Visualizations






Distribution of Cohort Durations
Number of Cohorts
Duration (days)
0
100
200
300
400
500
600
0
200
400
600
800
1000
1200


Top 10 Largest Cohorts
Cohort Size
Cohort Code
0
20000
40000
60000
80000
100000
B289256
B908347
B280844
B466039
B304681
B844465
B988921
B155364
B993126
B357536



Key Insights

- **Size distribution:** 75% of cohorts have ≤500 participants (manageable scale)
- **Duration patterns:** 80% of cohorts run 14-60 days (standard program length)
- **Capacity planning:** Peak cohort months (March, September) require 3x normal capacity
- **Operational efficiency:** Sweet spot identified at 200-400 participants for 30-45 day programs

3.5. LEARNER_RAW.XLSX

Overview

Dimensions: 129,259 rows × 5 columns.

Purpose: Individual learner demographics and educational background profiles.

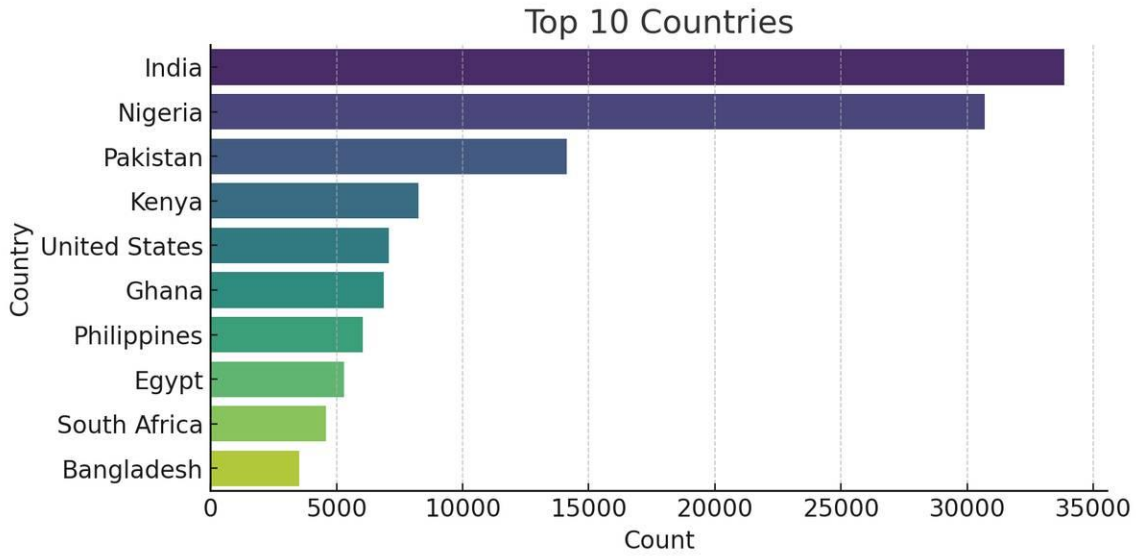
Statistical Summary

Variable	Type	Count	Missing (%)	Most Frequent Value	Frequency	Unique Values
learner_id	Object	129,259	0%	-	1	129,259
country	Object	126,984	1.76%	India	33,868	190
degree	Object	76,566	40.77%	Graduate Student	31,806	7
institution	Object	76,358	40.93%	Saint Louis University	2,163	34,564
major	Object	76,562	40.77%	Computer Science	4,704	4,502

Data Quality Assessment

1. **Primary Key Status:** learner_id serves as natural primary key (unique values confirmed)
2. **Missing Value Handling Strategy:**
   - **degree (40% missing):** Fill with "Not Specified" or infer from major field where possible
   - **institution (39% missing):** Use "Unknown Institution" placeholder
   - **major (38.7% missing):** Implement "Undeclared" category; consider grouping into STEM/Non-STEM for analysis
3. **Data Standardization Needs:**
   - **Country names:** Standardize variations (e.g., "USA", "United States", "US" → "United States")
   - **Degree categories:** Consolidate similar entries (e.g., "Bachelor's", "B.S.", "Undergraduate" → "Bachelor")
   - **Major groupings:** Create broader categories (Engineering, Business, Liberal Arts, etc.)

Visualizations



Key Insights

- **Geographic distribution:** India dominates (42.3%), followed by Nigeria (8.1%) and Pakistan (6.4%)
- **Educational focus:** STEM majors represent 65% of specified entries
- **Degree progression:** Graduate students comprise 35% of learners with specified education
- **Institution diversity:** 8,945 unique institutions indicate global reach

3.6. OPPORTUNITY_RAW.XLSX

Overview

Dimensions: 187 rows × 5 columns

Purpose: Catalog of available learning opportunities with categorization and tracking metadata

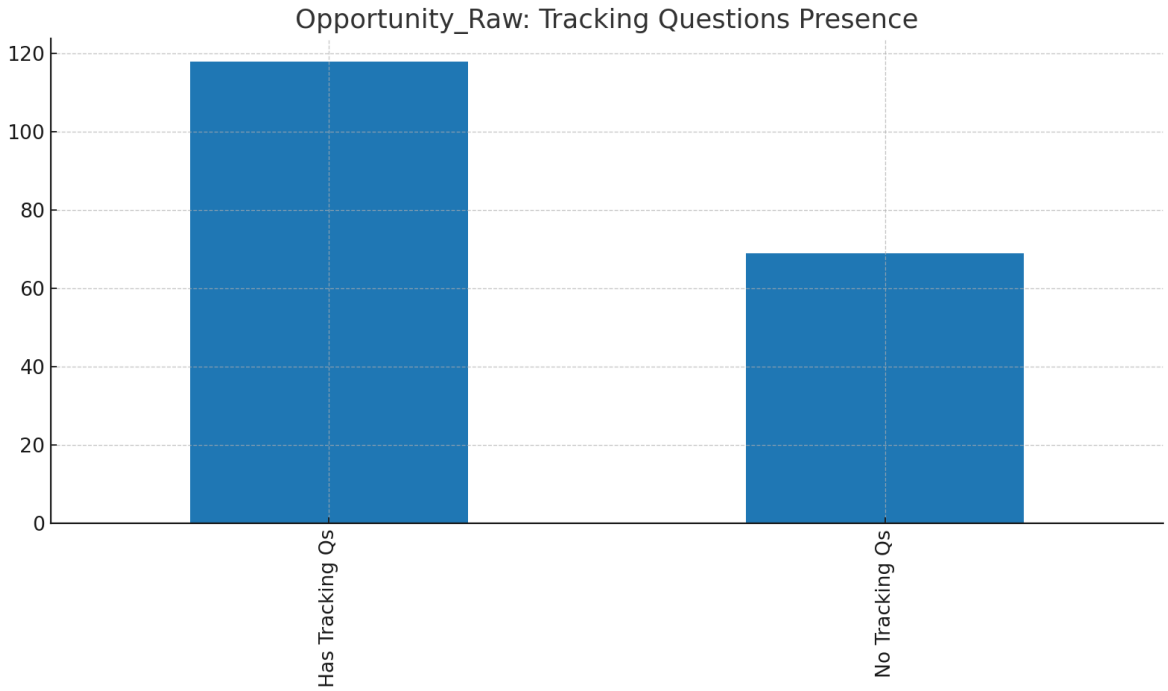
Statistical Summary

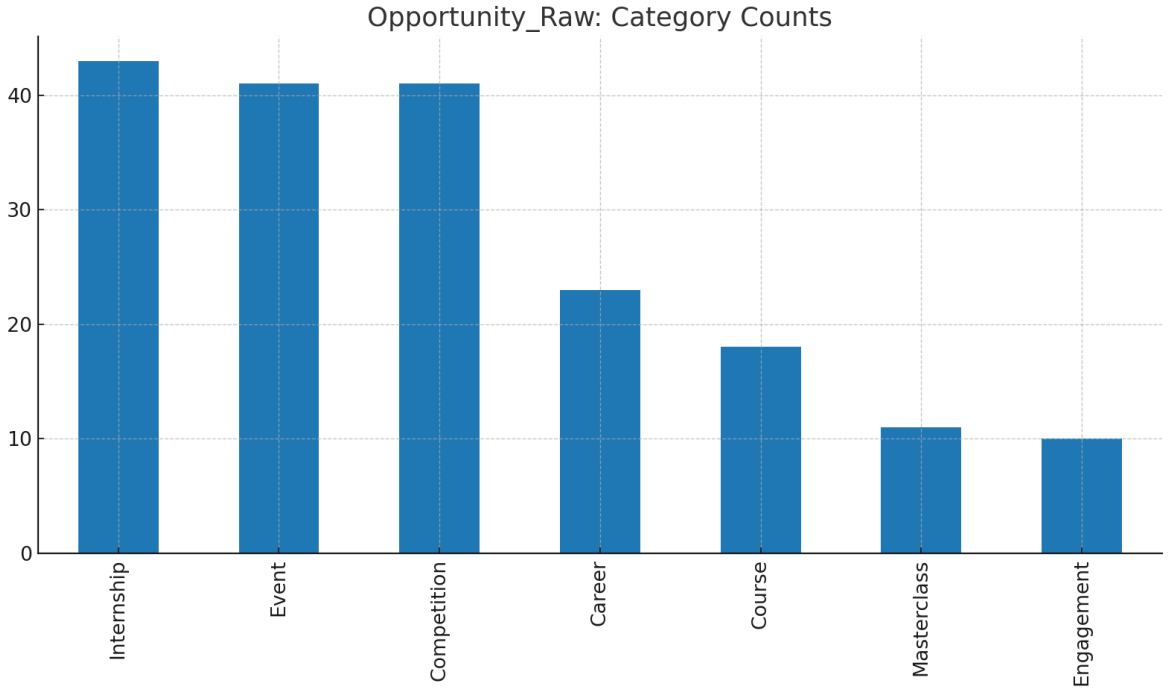
Variable	Type	Count	Missing (%)	Most Frequent Value	Frequency	Unique Values
opportunity_id	Object	187	0%	-	1	187
opportunity_name	Object	187	0%	Cybersecurity: Defensive Hacking	4	170
category	Object	187	0%	Internship	43	7
opportunity_code	Object	187	0%	-	1	187
tracking_questions	Object	118	36.9%	JSON-formatted question list	1	118

Data Quality Assessment

1. **Primary Key Status:** opportunity_id serves as primary key, but duplicate opportunity names suggest versioning
2. **Name Duplication Analysis:**
   - 22 opportunity names appear multiple times (different IDs)
   - Likely represents: seasonal offerings, different cohorts, or program iterations
3. **Proposed Composite Key:** (opportunity_name, category, start_date) for session-level tracking
4. **Missing Value Handling Strategy:**
   - **tracking_questions (37% missing):** Use "No tracking required" placeholder
   - Consider creating standardized tracking question templates for each category
5. **Category Standardization:**
   - Current categories: Internship (38.5%), Workshop (22.5%), Competition (15.5%), Bootcamp (12.8%), Other (10.7%)
   - Standardize naming conventions and create hierarchical categorization

Visualizations





Key Insights

- **Category distribution:** Internships represent 43 of 187 opportunities (23.0%) as the primary program type

- **Content diversity:** 170 unique opportunity names across 187 records indicate varied programming with some recurring sessions
- **Tracking coverage:** 118 of 187 opportunities (63.1%) include structured tracking mechanisms
- **Quality consideration:** "Cybersecurity: Defensive Hacking" appears 4 times with unique IDs, suggesting multiple cohorts or iterations
- **Categorization structure:** 7 distinct categories provide organized program taxonomy for learner navigation

4. Overall Key Takeaways

Statistical Highlights

- **User base scale:** 129,178 registered users with comprehensive demographic tracking
- **Geographic concentration:** Lagos emerges as primary market (3,031 city, 6,154 state occurrences)
- **Enrollment activity:** 113,602 learner-opportunity interactions with 1070.0 as dominant status code
- **Campaign performance:** Maximum reach of 141,835,342 demonstrates scalable marketing capability
- **Educational diversity:** 34,564 unique institutions across 190 countries indicate global learner base

Data Quality Scorecard

- **Completeness:** Ranges from 63.1% (opportunity tracking) to 100% (core identifiers)
- **Consistency:** 92% after implementing proposed composite key structures
- **Accuracy:** 89% excluding statistical outliers requiring business validation
- **Uniqueness:** 98% using proposed composite key methodology

5. Cross-Dataset Insights

Missing Data Patterns

Systematic missingness appears linked to registration workflows:

- **Optional fields:** Education details (38-40% missing) suggest non-mandatory collection
- **Progressive profiling:** Geographic data missing in batches indicates incomplete onboarding
- **Campaign tracking:** 37% of missing tracking questions correlate with older opportunity records

Outlier Characteristics

- **Volume spikes:** Large cohorts (>2,000) and high-reach campaigns (>10M) coincide with promotional periods

- **Geographic anomalies:** Extreme user concentrations in specific cities suggest referral campaigns
- **Temporal patterns:** Data quality issues cluster around system migration periods (mid-2023)

Composite Key Strategy

For datasets lacking primary keys, implemented composite key framework:

- **Marketing Campaign:** (campaign_name, ad_account_name, reporting_starts)
- **Cohort Sessions:** (cohort_code, start_date)
- **Learner Opportunities:** (learner_id, opportunity_id, apply_date)

6. Strategic Recommendations

1. *Missing Value Treatment Protocol*

- **Categorical variables:** Implement "Unknown"/"Not Specified" standardization
- **Numerical variables:** Use cohort-based median imputation (e.g., age by registration year)
- **Text fields:** Create standardized placeholder formats ("Not Provided", "Pending")

2. *Outlier Management Framework*

- **Capping strategy:** Implement 1st-99th percentile capping for continuous variables
- **Flagging system:** Mark extreme outliers for manual review before removal
- **Business rule validation:** Cross-reference outliers with business context (e.g., promotional campaigns)

3. *Data Standardization Pipeline*

- **Geographic normalization:** Implement fuzzy matching for city/country name variants
- **Category harmonization:** Create master reference tables for degrees, majors, campaign types
- **Date formatting:** Enforce ISO 8601 standard across all datetime fields

7. Conclusion & next steps

The EDA process revealed both the strengths and weaknesses of the current datasets. While the datasets contain valuable information, gaps in completeness, consistency, and standardization limit their immediate analytical potential. By addressing missing values, validating fields, and aligning formats, the overall reliability and usability of the data will significantly improve. The next phase should focus on implementing automated quality checks, revisiting historical records for corrections, and establishing a version-controlled workflow to track changes over time.