Exploratory Data Analysis (EDA) Report

Submitted by : 0707 DVA | Team 44
Date of Submission: 14-07-2025

1. Introduction

This report synthesizes the key findings, insights, and recommendations from the Exploratory Data Analysis performed on Six distinct datasets: Marketing Campaign Data, Learner Opportunity Data, Cognito Data, Cohort Data, Opportunity Data, and User Data. The aim is to provide a holistic view of the data quality, structure, and initial patterns observed across these different data sources.

2. Dataset Overviews

2.1. Learner Opportunity Dataset

- **Dataset Name/Purpose**: Contains information about learners' participation in different opportunities.
- **Number of Records and Columns**: 113,602 records and 5 columns.
- **Key Columns**:
    - enrollment_id: Unique identifier for each enrollment.
    - learner_id: Identifier for each learner.
    - assigned_cohort: Indicates the cohort assigned to the enrollment.
    - apply_date: Timestamp of the application (converted to datetime for analysis).
    - status: A numerical code representing the status of the enrollment.
- **Inferred Data Source**: Implies a system tracking learner opportunities.

2.2. Cognito Dataset

- **Dataset Name/Purpose**: User authentication and profile metadata, likely from Amazon Cognito. Used for account registration and login tracking.
- **Number of Records and Columns**: 129,178 rows and originally 17 columns, reduced to 9 meaningful columns after removing empty ones.

- **Key Columns**: user_id, email, gender, UserCreateDate, UserLastModifiedDate, birthdate, city, zip, state.
- **Inferred Data Source**: Cloud-based user authentication system (likely Amazon Cognito).
- **Type**: Historical snapshot.

2.3. User Dataset

- **Dataset Name/Purpose**: Explores the learner dataset to uncover patterns in academic profiles, institutional representation, and geographic diversity.
- **Number of Records and Columns**: Approximately 100 rows (verified record count post-import) and 12 columns.
- **Key Columns**: Learner_ID, Name, Age, Gender, Email, Phone, City, Start Date, Completion Date, Course_Name, Assessment_Score, Completion_Status.
- **Inferred Data Source**: Likely collected from a Learning Management System (LMS).
- **Type**: Historical, as it contains learner attributes and performance outcomes.

2.4. Marketing Campaign Dataset

- **Dataset Name/Purpose**: Inspects dataset quality, understands underlying structure, identifies key relationships between campaign attributes, and analyzes their impact on campaign performance metrics such as Results and Cost per result.
- **Number of Records and Columns**: 148 rows and 13 columns initially; 137 rows and 13 columns after dropping missing values.
- **Key Columns**: Ad Account Name, Campaign name, Delivery status, Delivery level, Reach, Outbound clicks, Landing page views, Result type, Results, Cost per result, Amount spent (AED), CPC (cost per link click), Reporting starts.
- **Inferred Data Source**: Marketing campaign platform.

2.5. Opportunity Dataset

- **Dataset Name/Purpose**: Contains opportunity records from Excelerate programs, detailing events, internships, and competitions offered to users. Supports analysis of program offerings and user engagement.
- **Number of Records and Columns**: 187 rows and 5 columns.

- **Key Columns**: opportunity_id, opportunity_name, category, opportunity_code, tracking_questions.
- **Inferred Data Source**: Likely CRM or system export (e.g., Salesforce, internal DB).
- **Type**: Historical/static.

2.6. Cohort Data Set

- **Dataset Name/Purpose**: Contains information about cohorts, likely from a learning/cohort management system.
- **Number of Records and Columns**: 640 records and 5 columns.
- **Key Columns**:
    - cohort_id: Placeholder value ("Cohort#") for all entries.
    - cohort_code: Unique code per cohort; acts as the Primary Key.
    - start_date: UNIX epoch timestamp (needs conversion to datetime).
    - end_date: UNIX epoch timestamp (needs conversion to datetime).
    - size: Integer, indicating cohort size; contains unusually large values.
- **Inferred Data Source**: Likely a CRM or database export from a learning/cohort management system.
- **Type**: One-time snapshot (historical).

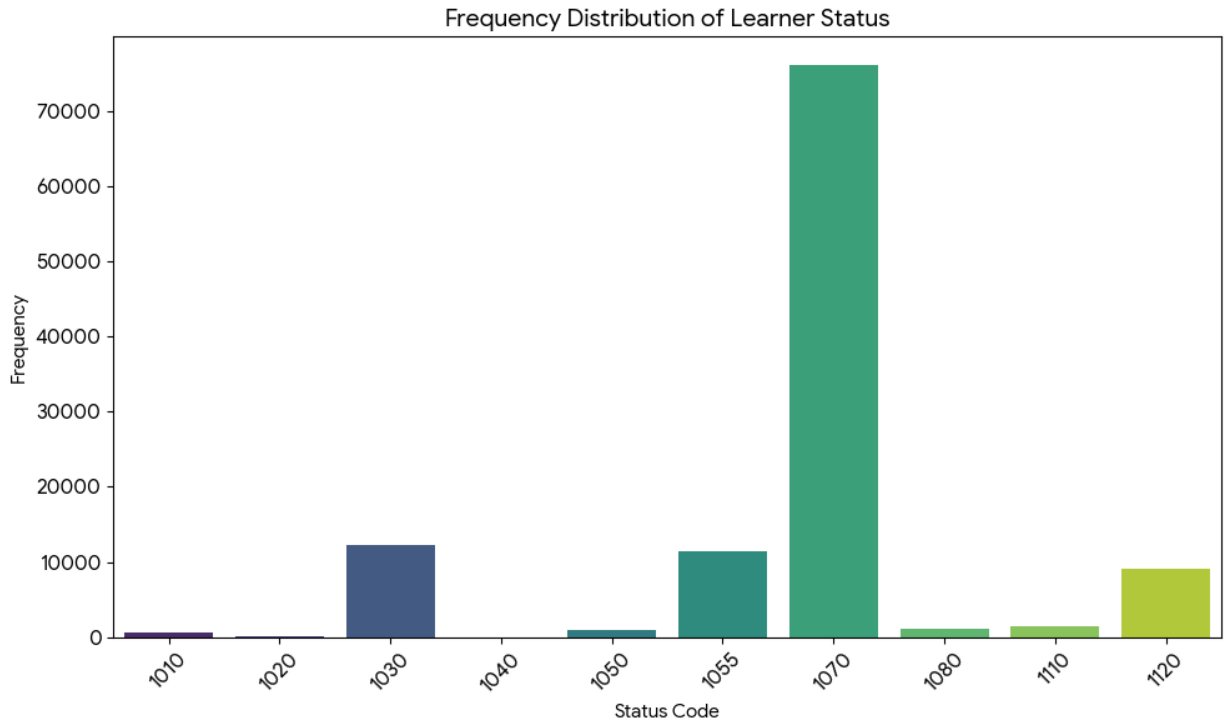
3. Summary of Key Findings (by Dataset)

3.1. Learner Opportunity Dataset

- **Missing Values**

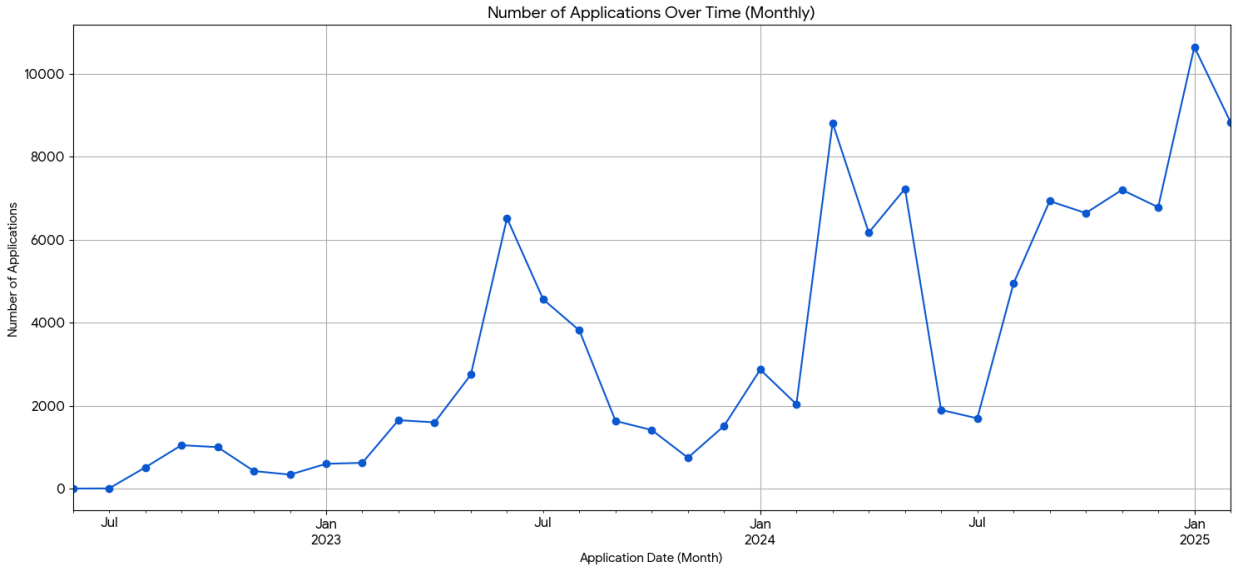
Column	Missing Count	Missing Percentage (%)
enrollment_id	0	0.0
learner_id	0	0.0
assigned_cohort	13,318	11.72
apply_date	188	0.17
status	186	0.16

- **Status Distribution**: Status 1070 is the most frequent outcome by a significant margin. Other notable statuses in terms of frequency include Status 1030, 1055, and 1120, while statuses such as 1010, 1020, 1040, 1050, 1080, and 1110 occur much less frequently. This distribution highlights that most learner opportunities progress to or end in Status 1070.



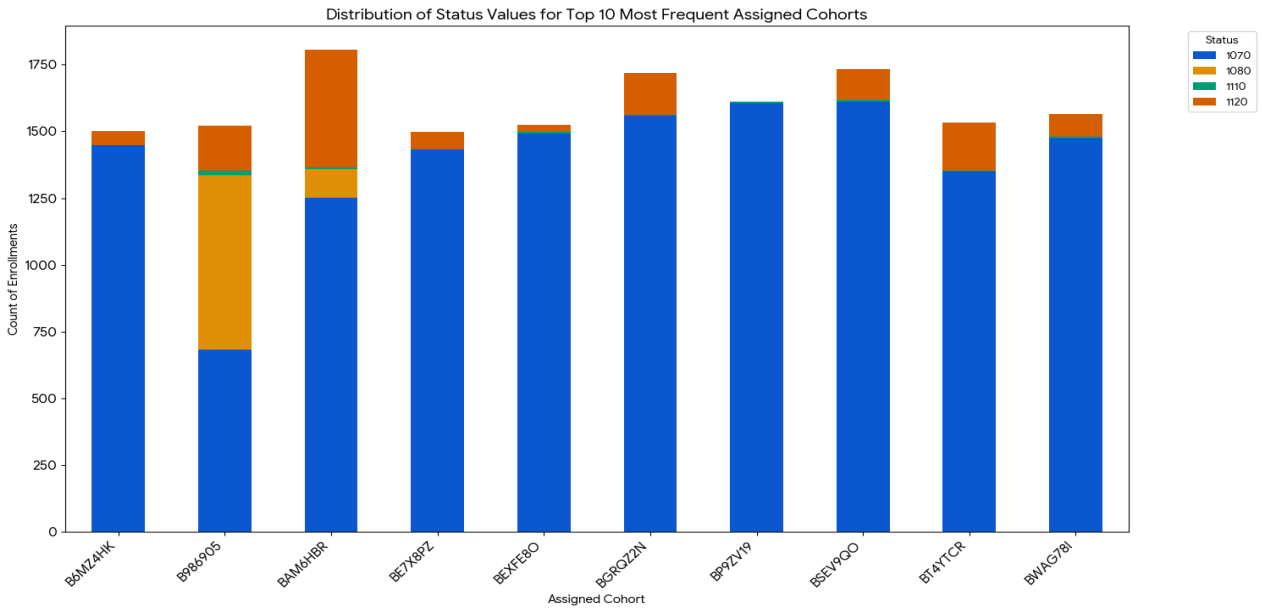
  - **Visualization**: *Frequency Distribution of Learner Status (Bar Chart)* - This chart visually confirms that Status 1070 dominates the distribution.

- **Application Trends**: The number of applications shows an upward trend over time, with significant peaks in June 2023, March and May 2024, and the highest peaks in January and February 2025. Conversely, there are periods with lower application volumes, such as the initial months in mid-2022 and towards the end of 2023. These trends can be valuable for understanding the seasonality and growth of learner opportunities.



- **Visualization**: *Number of Applications Over Time (Monthly Line Chart)* - This chart clearly illustrates the monthly application volume and highlights the identified peak periods.

- **Cohort-Status Relationship**: The distribution of statuses varies across different assigned_cohorts, with Status 1070 being predominant. Status 1080 is notably high for B986905, suggesting a unique outcome pattern for that specific cohort compared to others. Status 1120 is also significantly present in several cohorts.



- **Visualization**: *Distribution of Status Values for Top 10 Most Frequent Assigned Cohorts (Stacked Bar Chart)* - This chart visually represents how

the proportion of each status category varies across different cohorts, making it easy to identify cohorts with distinct outcome patterns.

- **Data Quality**:
    - assigned_cohort has a substantial amount of missing data (11.72%).
    - apply_date and status have a small percentage of missing values (less than 0.2% each).
    - No fully duplicated rows were found.
    - Outlier detection for status (using IQR) considers any status other than 1070 as an outlier due to the heavy concentration around 1070. This suggests that status might be more of an ordinal or categorical variable where "outliers" are simply less frequent categories.

3.2. Cognito Dataset

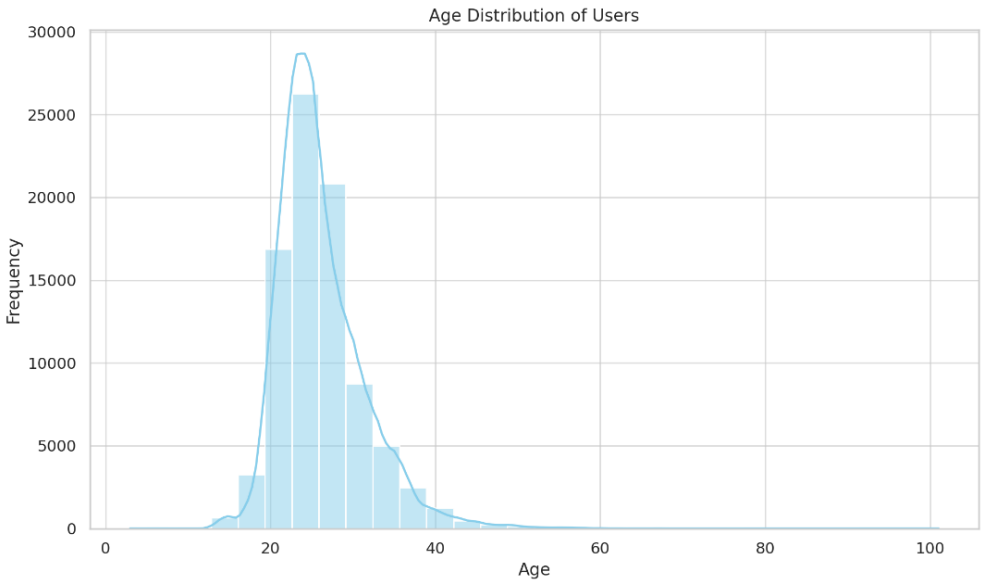
- **Data Structure**: The initial dataset contained 17 columns, 8 of which were entirely empty (Unnamed: 9 to Unnamed: 16) and were removed, leaving 9 meaningful columns (user_id, email, gender, UserCreateDate, UserLastModifiedDate, birthdate, city, zip, state).

- **Missing Values**:

Column	Missing Values
gender	42,862 missing (~33.2%)
birthdate	42,862 missing (~33.2%)
City	42,866 missing (~33.2%)
zip	42,870 missing (~33.2%)
state	42,937 missing (~33.2%)

- **Duplicate Records**: A full scan found 0 duplicate rows; user_id uniquely identifies each user and acts as a Primary Key.
- **Age Distribution & Outliers**:
    - A new age column was derived from birthdate.
    - The user base is predominantly young adults, with most ages falling between 23–29.

- Age statistics: Count: 86,316; Mean: 26.1; Median: 25; Std. Dev: 5.27; Min: 3; Max: 101.
- Using the IQR method, 2,288 user records were identified as age outliers (below 14 or above 38 years). These include unusually young (3–13 years) and older (39–101 years) users, which may indicate data entry errors or test data.



- **Visualization**: *Age Distribution of Users (Histogram with KDE)* - This plot visually confirms the concentration of users in the 20-30 age range and highlights the presence of outliers at the extremes.

- **Data Quality Issues Identified**:
  - **Missing Data**: High percentage (~33%) of missing values in multiple columns (gender, birthdate, city, zip, state).
  - **Text Dates**: UserCreateDate and UserLastModifiedDate are stored as plain text and require conversion to proper datetime format.
  - **Outliers**: Implausible ages (e.g., 3 or 101 years) indicate potential anomalies.
  - **Empty Columns**: Unnamed: 9 to Unnamed: 16 were entirely empty and removed.
  - **Text Inconsistency**: Fields like gender, city, and state may have inconsistent capitalization and formatting (e.g., "NAIROBI", "nairobi") and should be standardized.

3.3. User Dataset

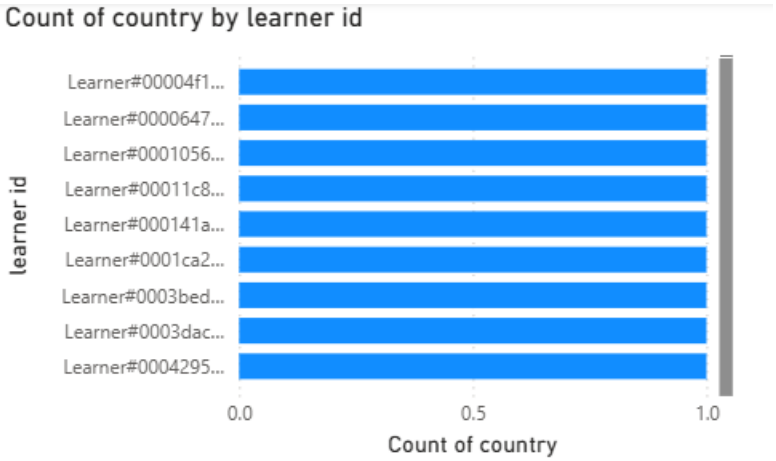
- **Dataset Purpose**: To uncover patterns in academic profiles, institutional representation, and geographic diversity.
- **Column Meanings & Data Types**: Contains categorical attributes such as Learner_ID, Country, Degree, Institution, and Major, along with Name, Age, Gender, Email, Phone, Start Date, Completion Date, Course_Name, Assessment_Score, and Completion_Status.
- **Data Quality Issues**:

Column Name	Description (Assumed)	Data Type	Issues Noted	Missing values
Learner_ID	Unique identifier for each learner	Text	Looks consistent	0
Name	Learner's full name	Text	OK	0
Age	Age in years	Numeric	Valid but missing values	6
Gender	Gender	Categorical	Case inconsistencies (e.g., "Male", "male")	0
Email	Learner email address	Text	OK	0
Phone	Contact number	Text	Formatting varies	1
City	Learner's city	Categorical	OK	0
Start_Date	Program start date	Date (Text Format)	Needs conversion to date	0

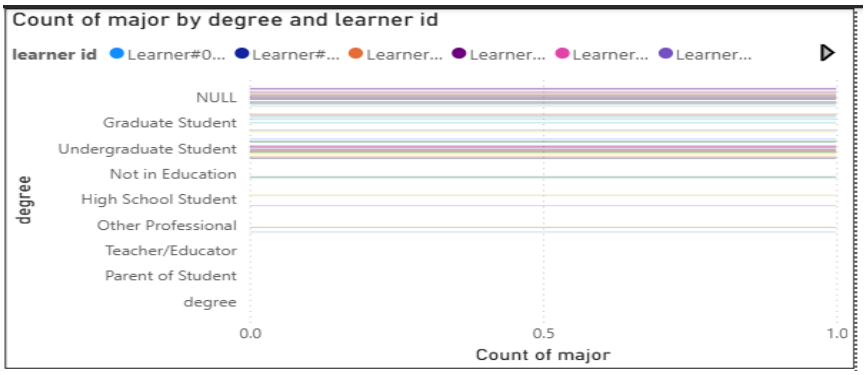
Completion_ Date	Program completion date	Date (Text Format)	Needs conversion to date	12
Course_Name	Name of enrolled course	Text	OK	0
Assessment _Score	Score achieved	Numeric	Some outliers, missing values	9
Completion_ Status	Course completion status	Categorical	Some Inconsistencies in values (e.g., "Completed", "completed")	2

  - Duplicate institutions may exist with slight name variations.
  - No critical errors or mismatches found post-import to PostgreSQL.

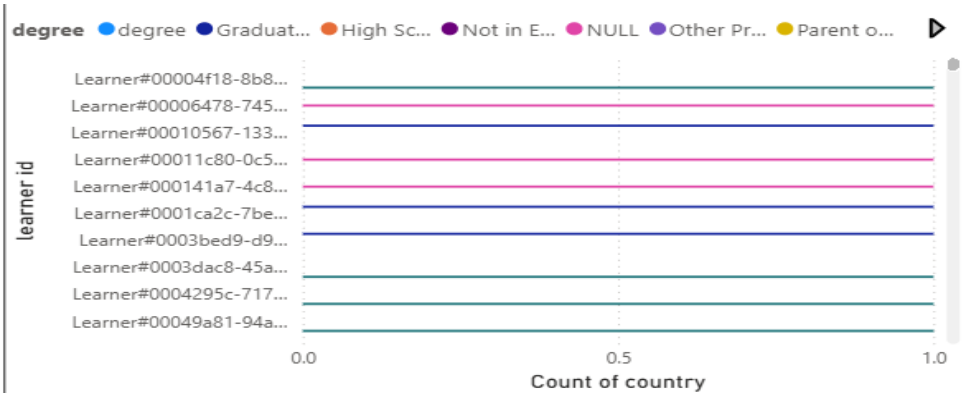
- **Key Relationships**: Learner_ID is a Primary Key. Course_Name could be used as a Foreign Key. Completion_Status is used for outcome analysis.
- **Insights from Visualizations (Power BI Dashboards)**:
  - **Learner Distribution by Country**: Identified dominant regions (e.g., India, USA) contributing the most learners.



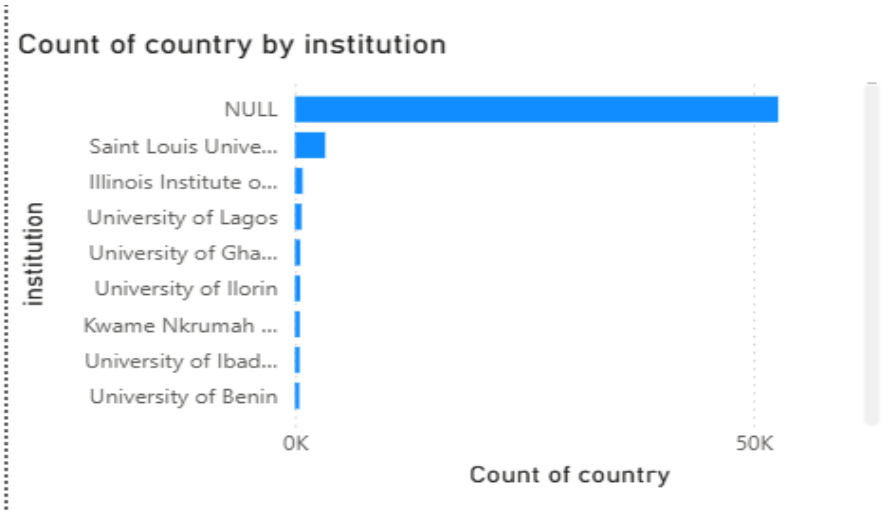
- **Degree Popularity by Major**: Technical degrees were mostly paired with majors like Computer Science and Engineering, while business degrees leaned toward Marketing and Economics.



- **Institution-wise Major Spread**: Highlighted schools with strong specialization (e.g., a particular college hosting most Data Science learners).



- **Distinct Institutions by Country**: Revealed countries with greater academic diversity in the dataset.



- **Relationship Analysis**:
  - **Country-Major**: Some majors appear heavily in specific countries (e.g., Engineering in India, Business in the US), indicating regional preferences or education system focus.
  - **Degree-Major**: Degrees are aligned predictably with majors (e.g., B. Tech with technical fields), suggesting consistent academic pathways.
  - **Institution-Major**: Certain majors were concentrated in just one or two institutions, highlighting specialization or flagship programs.
  - **Country-Institution Diversity**: Some countries have learners from many institutions, while others are concentrated, reflecting centralization vs. distribution of academic opportunities.

3.4. Marketing Campaign Data (2023-2024) Dataset

- **Data Quality**: The dataset was relatively clean after addressing a small percentage of missing values (ranging from 4.05% to 6.08%). Rows with any missing values were removed, resulting in a cleaned dataset of 137 rows.

Missing Values Summary:

Column Name	Missing Count	Missing Percentage
Ad Account Name	7	4.73%
Campaign name	9	6.08%
Delivery status	7	4.73%
Delivery level	7	4.73%
Reach	7	4.73%
Outbound clicks	9	6.08%
Landing page views	9	6.08%
Result type	7	4.73%
Results	6	4.05%
Cost per result	7	4.73%
Amount spent (AED)	6	4.05%

CPC (cost per link click)	8	5.41%
Reporting starts	7	4.73%

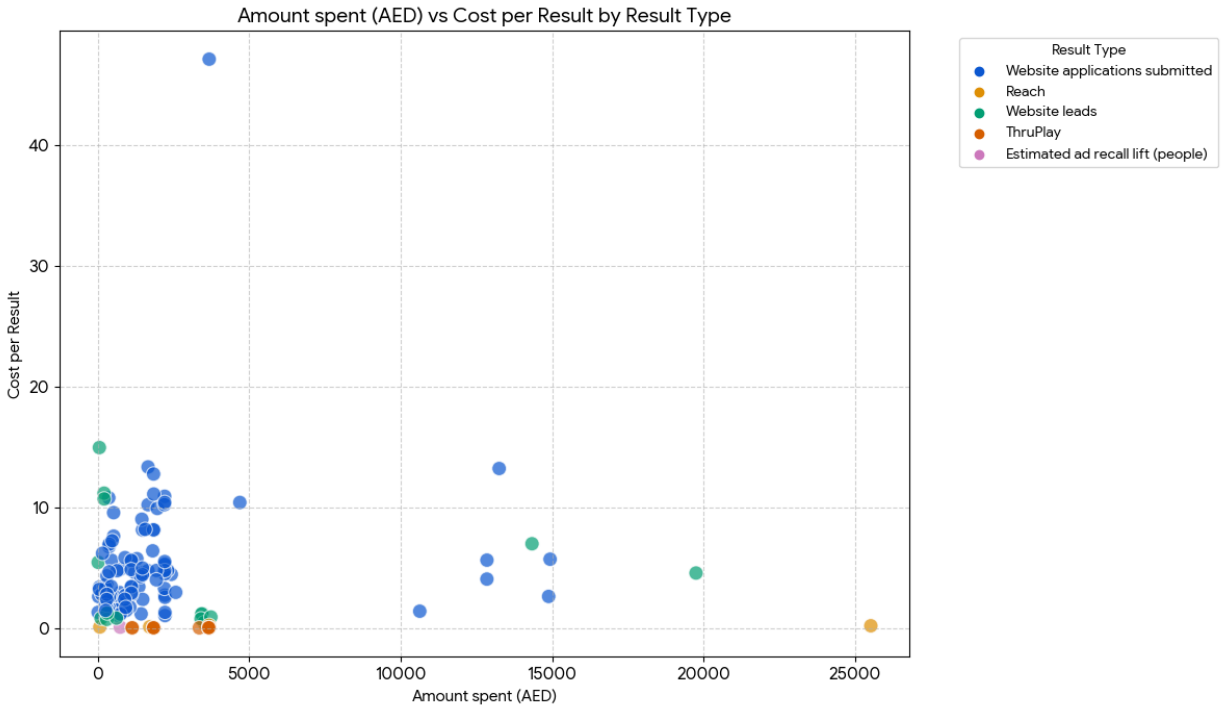
Missing Data Handling:

Given that the percentage of missing values in each column was relatively low (all below 7%), the chosen handling method was to remove rows with any missing values. This approach was selected to maintain data integrity and avoid introducing potential biases that might arise from imputation methods, especially considering the mixed data types (numerical and categorical) and the varied nature of campaign data. After dropping rows with missing values, the cleaned dataset consists of 137 rows and 13 columns.

- **Numerical Column Statistics**:
    - Reach and Results exhibit extremely high maximum values and large ranges, indicating a few campaigns with exceptionally high performance, leading to highly skewed distributions (mean much higher than median).
    - Outbound clicks and Landing page views also show right-skewed distributions.
    - Cost per result and CPC (cost per link click) have relatively smaller ranges, suggesting more consistent cost performance across campaigns.
- **Correlations (Pearson)**:
    - **Strong Relationship between Reach and Results**: An almost perfect positive correlation (0.998) indicates that Reach is a direct and significant determinant of Results.
    - **Amount spent (AED) drives Outbound clicks**: A very strong positive correlation (0.88) suggests that higher spending effectively increases the number of outbound clicks.
    - **Spending and Reach/Results**: Amount spent (AED) shows moderate positive correlations with Reach (0.55) and Results (0.52), implying that

increased investment generally leads to broader reach and more overall results.

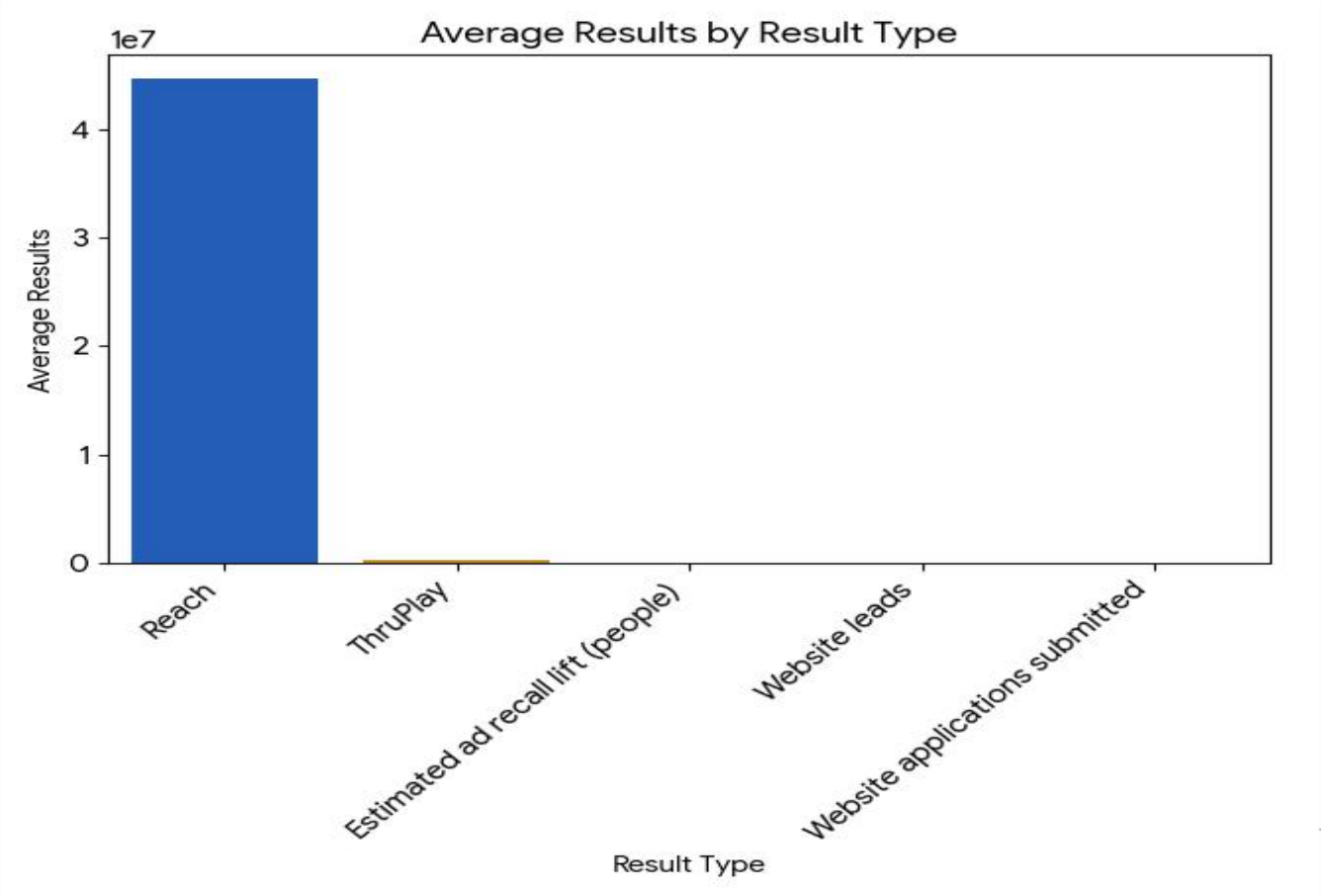
- **Cost per result independence**: Cost per result has weak or negligible linear correlations with other numerical metrics, suggesting it's influenced by factors beyond simple linear relationships with Reach, Amount spent, or Outbound clicks.

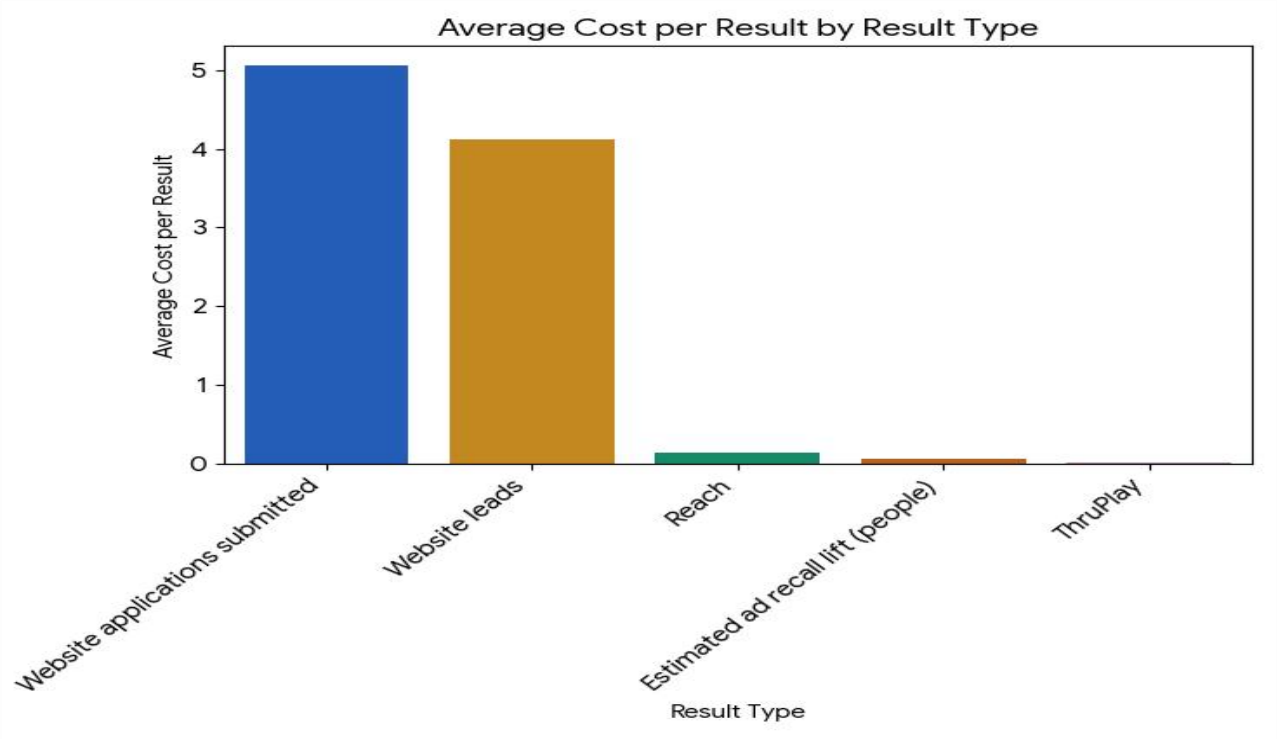


- **Visualization**: *Amount spent (AED) vs Cost per Result by Result Type (Scatter Plot)* - This plot visually confirms the positive relationships, albeit with some dispersion due to outliers.

- **Impact of Result type on Performance**:
  - **Objective-Dependent Performance**: The Result type significantly dictates both the volume of Results and the Cost per result.
  - **High Volume, Low Cost for Awareness**: Campaigns aimed at Reach and ThruPlay generate a massive volume of Results at a very low Cost per result (e.g., $0.14 for Reach, $0.01 for ThruPlay).
  - **Higher Cost for Conversions**: Conversion-focused objectives like Website applications submitted and Website leads naturally incur a much higher

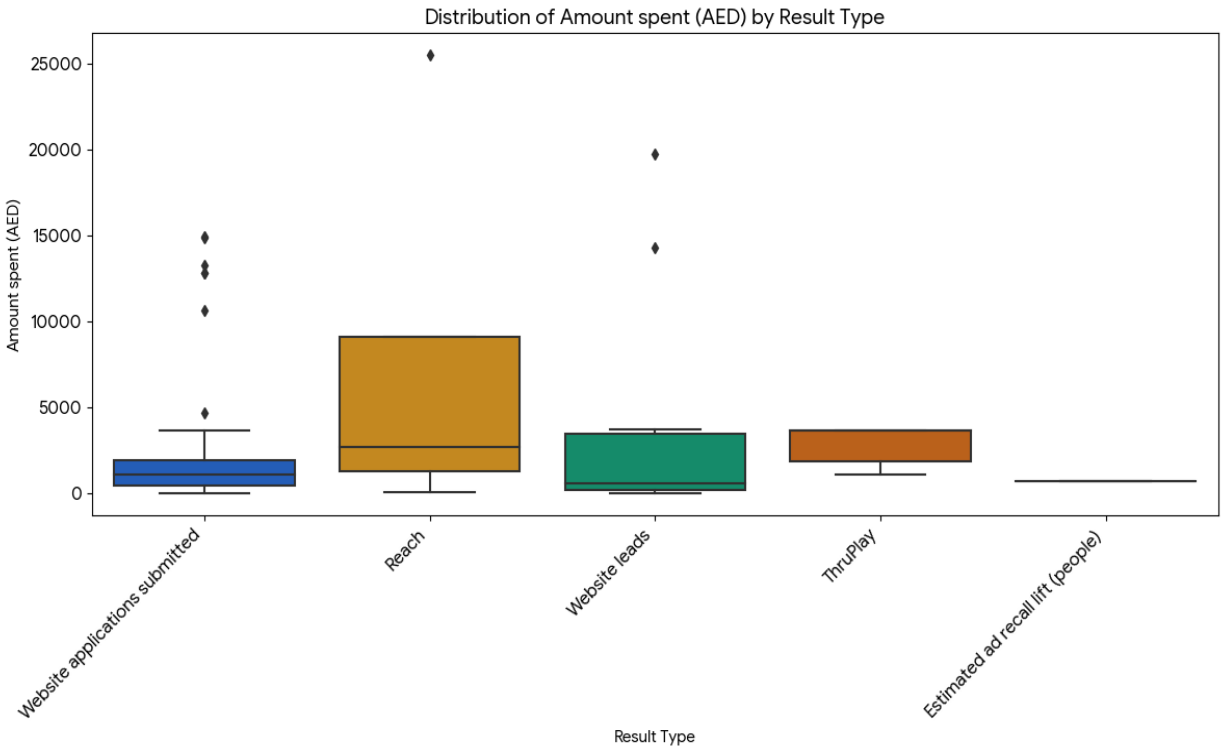
Cost per result ($5.05 and $4.11 respectively), reflecting their position at the bottom of the marketing funnel.





- **Visualizations**: *Average Results by Result Type (Bar Plot)* and *Average Cost per Result by Result Type (Bar Plot)* - These plots clearly illustrate these differences.

- **Spending Distribution by Result Type**: Campaigns focused on Reach tend to have the highest average and maximum spending, suggesting they are often allocated larger budgets. Website leads and Website applications submitted also show significant variability in spending.

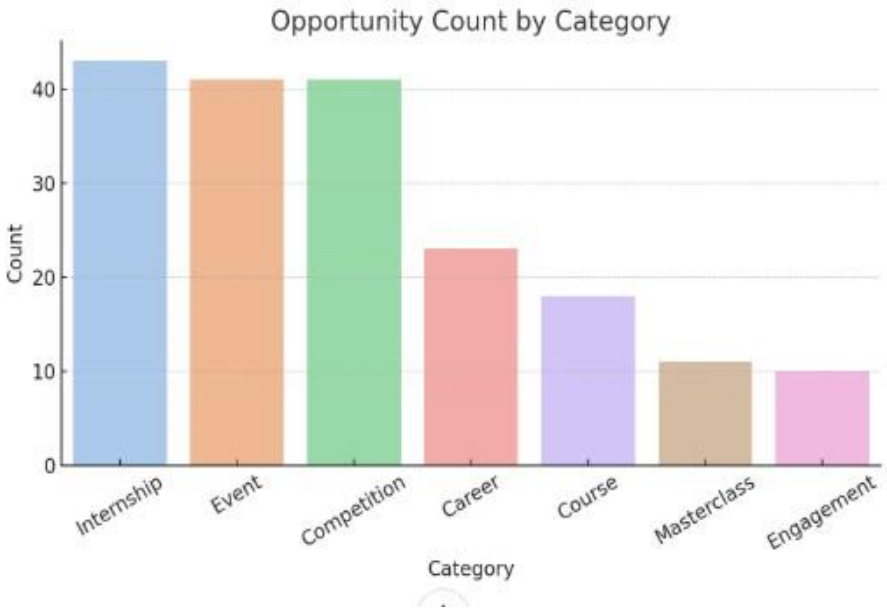


- **Visualization**: *Distribution of Amount spent (AED) by Result Type (Box Plot)* - This plot illustrates the spending distribution across different result types.

- **Campaign Timing and Spending (Reporting starts)**: The data points for Reporting starts are very sparse and mostly concentrated around specific dates, suggesting it represents specific campaign launch or reporting dates rather than continuous daily reporting. This limitation makes it challenging to discern clear daily spending trends.

3.5. Opportunity Dataset

- **Dataset Size**: 187 rows and 5 columns.
- **Key Columns**: opportunity_id (Unique ID, Primary Key), opportunity_name (Text, not unique), category (Categorical: Event, Competition, etc.), opportunity_code (Appears unique, possible Secondary Key), tracking_questions (Contains JSON-like strings).
- **Category Distribution**: Internship and Event are the most frequent categories. Competition, Career, Course, Masterclass, and Engagement are less common.



- **Visualization**: *Opportunity Count by Category (Bar Chart)* - This plot visually confirms the distribution of opportunities across different categories.

- **Data Quality Issues**:
  - **Missing Values**: 69 entries (37%) are missing in tracking_questions.
  - **Inconsistencies**: tracking_questions contains JSON-like strings stored as raw text, needing decoding.
  - **Data Types**: All fields are initially stored as text (object) type.
  - **No Duplicates**: No exact row duplicates were found.
  - **Limitations**: Missing tracking data, inconsistent naming potential, and no timestamp fields available (unknown update frequency).

3.6. Cohort Data Set

- **Data Structure**: The dataset contains 640 records and 5 columns. cohort_code is the primary key and uniquely identifies each cohort. cohort_id contains the same placeholder value ("Cohort#") for every row, making it unusable as an identifier.
- **Missing Values:**
  - No obvious nulls or blanks were detected in the first previewed rows.

- That said, date columns (start_date, end_date) are stored as epoch timestamps, not readable date formats — while technically complete, they're not usable until converted.

- **Potential Outliers**:
  - The size column (indicating cohort size) contains suspiciously high values, with multiple entries listed as 100,000. These values are unusually large for a cohort and may represent a system default/fallback, a data entry error, or a valid but rare extreme case.
  - **IQR-based outlier detection on the size column**:
    - Q1 (25th percentile): 500
    - Q3 (75th percentile): 1500
    - IQR (Q3 - Q1): 1000
    - Lower Bound: -1000 (not applicable since size cannot be negative)
    - Upper Bound: 3000
  - **Outliers Detected**: 47 cohorts have a size greater than 3000, which are considered outliers based on IQR. These unusually large cohort sizes could distort averages and visualizations and should be flagged for further review or excluded depending on context.
- **Basic Statistics for size**:
  - Total Records (Count): 640
  - Mean (Average): 5,741
  - Median: 800
  - Minimum Size: 3
  - Maximum Size: 100,000
  - Standard Deviation: ~20,994
  - **Observation**: The mean is heavily skewed by very large values (like 100,000), while the median is much lower (800), indicating the presence of significant outliers.

- **Limitations**:
    - start_date and end_date are in UNIX epoch format and must be converted for usability.
    - All cohort_id entries are identical ("Cohort#"), making this column unusable.
    - Unusually large size values may be outliers and require further investigation.

4. Cross-Dataset Synthesis

Several common themes and data quality challenges emerge across these six datasets:

- **Missing Data**: A recurring issue, ranging from small percentages (Learner Opportunity, Marketing Campaign) to substantial portions (Cognito's demographic fields, Opportunity Records' tracking_questions). This necessitates robust missing data handling strategies. The Cohort Data Set, while not having explicit nulls, has functionally "missing" usability in its date columns due to their format.
- **Date Field Formatting**: Multiple datasets (Cognito, User Data, Learner Opportunity, Cohort Data) contain date fields stored as text or UNIX epoch timestamps, requiring consistent conversion to proper datetime objects for accurate temporal analysis.
- **Categorical Data Inconsistencies**: Inconsistent capitalization and formatting are present in categorical fields across Cognito, User Data, and Opportunity Records, highlighting the need for standardization and normalization. The cohort_id in the Cohort Data Set also represents a categorical inconsistency due to its uniform placeholder value.
- **Dominant Categories/Skewed Distributions**: The status column in the Learner Opportunity dataset and Reach/Results in the Marketing Campaign data show highly skewed distributions or dominant categories. The size column in the Cohort Data Set also exhibits a highly skewed distribution due to extreme outliers. These characteristics impact how traditional statistical measures (like outlier detection) should be interpreted.
- **Outliers**: Outliers were identified in numerical fields like age (Cognito), status (Learner Opportunity), and size (Cohort Data), as well as in performance metrics (Marketing Campaign). These require careful consideration for their impact on analysis.

- **Unique Identifiers**: Most datasets have clear primary keys (enrollment_id, user_id, Learner_ID, opportunity_id, cohort_code), which are crucial for potential future data integration. However, the cohort_id in the Cohort Data Set is an exception, being unusable as a unique identifier.

5. Overall Recommendations for Further Analysis/Action

Based on the consolidated EDA, the following recommendations are crucial for preparing these datasets for deeper analysis and predictive modeling:

1. **Comprehensive Data Cleaning and Preprocessing**:
    - **Missing Value Imputation/Handling**: Implement strategies (e.g., imputation, removal, or separate analysis) for columns with significant missing data across all datasets.
    - **Date/Time Conversion**: Systematically convert all date-related columns from text or epoch timestamps to appropriate datetime formats.
    - **Categorical Data Standardization**: Apply consistent casing (e.g., proper case, uppercase) and remove leading/trailing spaces for all categorical fields. Resolve duplicate entries with slight name variations (e.g., institutions, majors). Address the unusable cohort_id in the Cohort Data Set (e.g., remove or mark as irrelevant).
    - **Outlier Management**: Develop clear policies for handling identified outliers, considering their nature (e.g., data errors vs. rare but valid occurrences). Specifically, investigate the large size values in the Cohort Data Set.
2. **Feature Engineering**:
    - **Temporal Features**: Extract granular temporal features (e.g., year, month, day of week, hour) from date columns to enable more detailed time-based pattern analysis across datasets.
    - **Categorical Encoding**: Prepare categorical variables for modeling by applying appropriate encoding techniques (e.g., one-hot encoding, label encoding).
    - **JSON Decoding**: Decode and potentially flatten JSON-like strings in tracking_questions (Opportunity Records) into structured features or a new table.
3. **Advanced Relationship Analysis**:

- **Cohort Performance**: Deep dive into the relationship between assigned_cohort and status in the Learner Opportunity dataset to understand factors driving specific outcomes within cohorts.
- **Marketing Efficiency Drivers**: Further investigate factors beyond Amount spent (AED) that influence Cost per result in marketing campaigns, such as audience targeting, ad creatives, and optimization strategies.
- **Learner Engagement**: Analyze the patterns of top learners (learner_id with many enrollments) in the Learner Opportunity dataset to understand their engagement drivers.
- **Cohort Size Impact**: Explore the impact of size on cohort outcomes or other related metrics, considering the presence of outliers.

4. **Data Integration and Master Data Management**:
    - **Cross-Dataset Joins**: Explore opportunities to join datasets using common identifiers (e.g., user_id from Cognito with learner_id from Learner Opportunity if a mapping exists, and cohort_code from Cohort Data with assigned_cohort from Learner Opportunity) to create richer profiles and enable more comprehensive analysis.
    - **Master Tables**: For datasets like User Data, build dimension tables for key entities (Country, Degree, Institution, Major) and consolidate cleaned data into a standardized master table.
5. **Improved Data Collection**:
    - **Temporal Granularity**: For future data collection, especially in marketing campaigns, aim for more granular and continuous Reporting starts dates to enable robust time-series analysis of spending and performance.
    - **Completeness**: Implement measures to reduce missing data in optional demographic fields during user registration or data entry.
    - **Data Export Quality**: Review export processes for datasets like Cohort Data to ensure all columns are meaningful and correctly formatted (e.g., cohort_id should contain unique, relevant values).

6. Conclusion

This consolidated EDA report provides a comprehensive overview of the six datasets, highlighting their individual characteristics, key findings, and common data quality challenges. By addressing these challenges through systematic cleaning, feature engineering, and targeted analyses, the organization can unlock deeper insights into learner behavior, marketing campaign effectiveness, user demographics, and cohort dynamics. This foundational work will significantly enhance the reliability and utility of these datasets for informed decision-making and the development of predictive models.