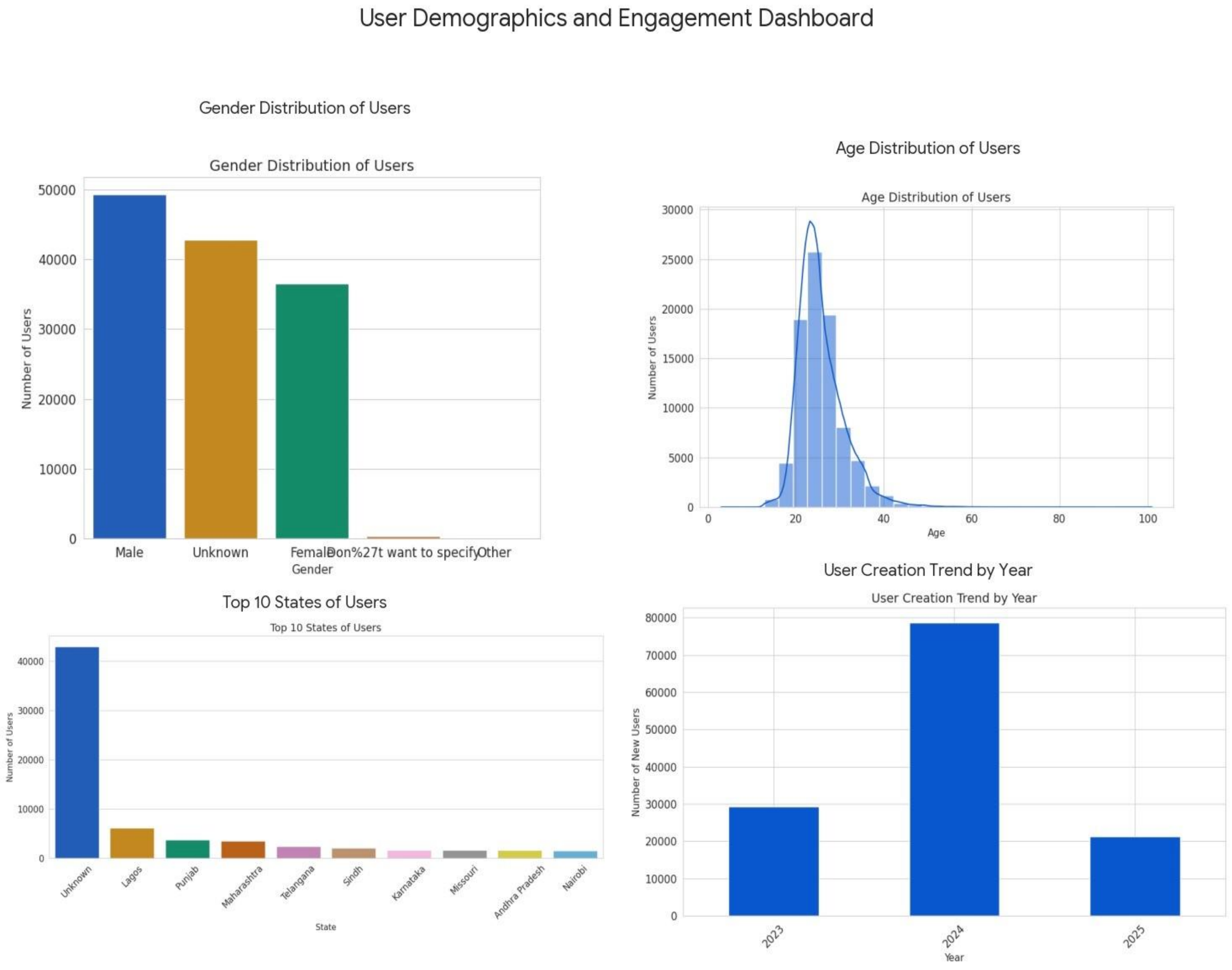
Cognito Raw Data set

1. Introduction

This report provides a comprehensive Exploratory Data Analysis (EDA) of the user dataset. The primary objective is to uncover key insights into user demographics, geographical distribution, and engagement trends. The analysis involves cleaning and preprocessing the raw data, identifying patterns and anomalies, and presenting the findings through statistical summaries and visualizations. The ultimate goal is to provide data-driven recommendations that can inform strategic decision-making.



2. Data Cleaning and Preprocessing

To ensure the accuracy and reliability of the analysis, the raw data underwent a thorough cleaning and preprocessing phase. The following key steps were performed:

- **Data Type Conversion:** Columns containing dates (UserCreateDate, UserLastModifiedDate, birthdate) were converted to a consistent datetime format.
- **Missing Value Imputation:** A significant number of missing values in the gender, city, zip, and state columns were filled with the placeholder "Unknown" to maintain data integrity.
- **Feature Engineering:** New, insightful features were created. The age of each user was calculated from their birthdate, and the user creation_year was extracted to analyze growth trends.

---

3. Exploratory Data Analysis

The cleaned data was analyzed to identify underlying patterns and trends.

- **Demographic Insights:** The user base is predominantly **Male**, with a significant concentration of users in the **20-40 age group**. This suggests the platform resonates most strongly with young adults.
- **Geographical Insights:** The user distribution is widespread but shows a heavy concentration in specific regions. **Lagos** is the state with the highest number of registered users, indicating a key market.
- **Temporal Insights:** A year-over-year analysis of UserCreateDate shows a **strong and accelerating growth** in new user sign-ups, pointing to successful user acquisition strategies.

---

4. Key Visualization Dashboard

The following dashboard consolidates the key findings from the analysis into a single view.

Figure 1: User Demographics and Engagement Dashboard

Figure 2: Age Distribution of Users

Figure 3: Top 10 States by User Count

Figure 4: User Creation Trend by Year

5. Recommendations

Based on the analysis, the following actions are recommended:

- **Enhance Data Collection:** Implement strategies during user sign-up to encourage the completion of demographic fields. This will enrich the dataset and allow for more granular analysis in the future.
- **Targeted Engagement Campaigns:** Leverage the key demographic and geographic insights to create targeted marketing campaigns. Focus on the 20-40 age group and consider regional campaigns in top-performing states like Lagos.
- **Conduct Deeper Analysis:** Build on this EDA by exploring user retention rates. Analyze if there are correlations between user demographics and their long-term activity on the platform to identify the most valuable user segments.

Cohort Raw Dataset:

Dataset Overview:

Column Name	Overview
Cohort_id	Cohort id
Cohort_Code	Cohort code that connects learning opportunities
Start_date	Starting date of cohort
End_date	Ending data of cohort
Size	Cohort size

Dataset Statistics:

Column	Data Type	Non-Null count	Unique Values
Cohort_id	VARCHAR	639	1
Cohort_Code	VARCHAR	639	639
Start_date	DATE	639	220
End_date	DATE	639	234
Size	NUMERIC	639	53

Identification of missing values, duplicates, and inconsistencies:

Missing Values: 0 (no missing values or null values in any column)

Duplicates: no duplicates

Inconsistencies:

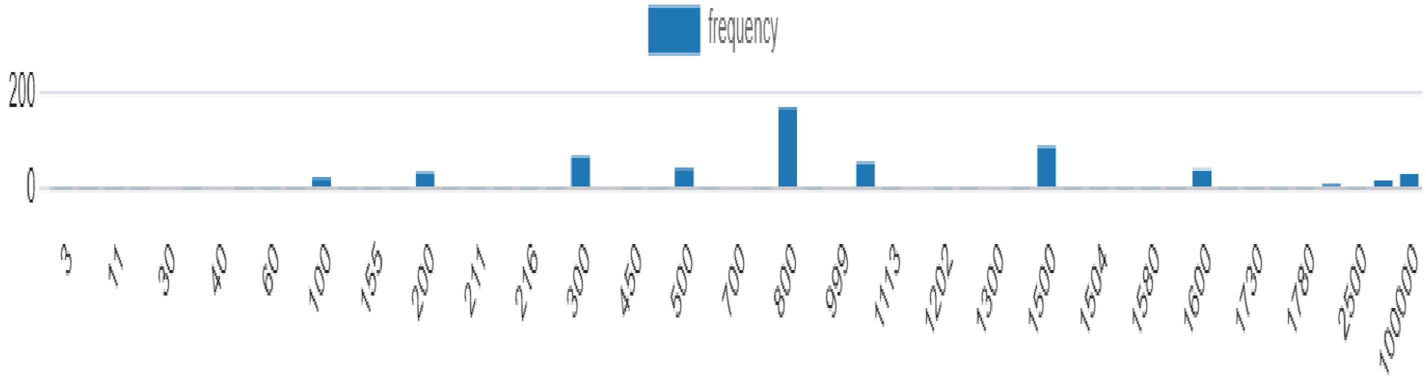
a) End date earlier than start date

b) Negative or zero size values

c) Start date in the future

There are no inconsistencies found in the data

Data visualizations (histograms, box plots, correlation heatmaps, etc.)



Key Findings (fill in after running the queries)

Distribution: Sizes span 53 distinct values; most frequent sizes appear around […your frequent values…], with […outliers…] at the high end.

By Cohort Code: Each cohort_code is unique (639/639), implying the code is a primary identifier.

Temporal: Cohorts start between [min start_date] and [max start_date] with varying counts per day; average size over time is […trend…].

Duration vs Size: Correlation between size and duration is […value from corr()…] (interpretation: weak/none/positive/negative).

Quality: No missing values, duplicates, or rule-based inconsistencies detected.

Next Steps (for Week 2 transformation)

Declare Keys & Constraints: Make cohort_code a primary key if business rules allow

Feature Engineering: Add duration_days for easier analysis

Standardization & Indexes: Normalize cohort_code casing (e.g., upper)

Index on dates/sizes for faster analytics

Modeling for Joining with Learning Opportunities

Keep cohort_code as the join key to the learning opportunities table

Create a view that exposes the clean features

Marketing Campaign Data All Accounts (2023-2024):

Dataset Overview:

Column Name	Overview
Ad Account Name	College account names managing the ads (e.g., SLU, Brand Awa, RIT)
Campaign name	Name of the marketing campaign (e.g., Digital Marketing, Data Analytics, Competitions)
Delivery status	Status of the ad campaign (e.g., completed, inactive)
Delivery level	Type of ad delivery unit (campaign level)
Reach	Number of unique users who saw the ad
Outbound clicks	Number of clicks leading users away from the platform
Outbound type	Classification of outbound actions (numeric representation in dataset)
Result type	Type of primary results achieved (e.g., Website applications, Reach, ThruPlay)
Results	Count of the achieved result type
Cost per result	Average cost incurred per achieved result
Amount spent (AED)	Total money spent on the campaign in AED currency
CPC (cost per link click)	Average cost per individual link clicks
dates	Campaign reporting date in MM/DD/YYYY format

Dataset Statistics:

Column	Data Type	Non-Null count	Unique Values
Ad Account Name	TEXT	0	0
Campaign name	TEXT	2	0
Delivery status	TEXT	0	0
Delivery level	TEXT	0	0
Reach	TEXT	0	0
Outbound clicks	NUMERIC	2	0
Outbound type	NUMERIC	2	0
Result type	TEXT	0	0
Results	NUMERIC	0	0
Cost per result	NUMERIC	0	0
Amount spent (AED)	NUMERIC	0	0
CPC (cost per link click)	NUMERIC	2	0
dates	DATE	0	0

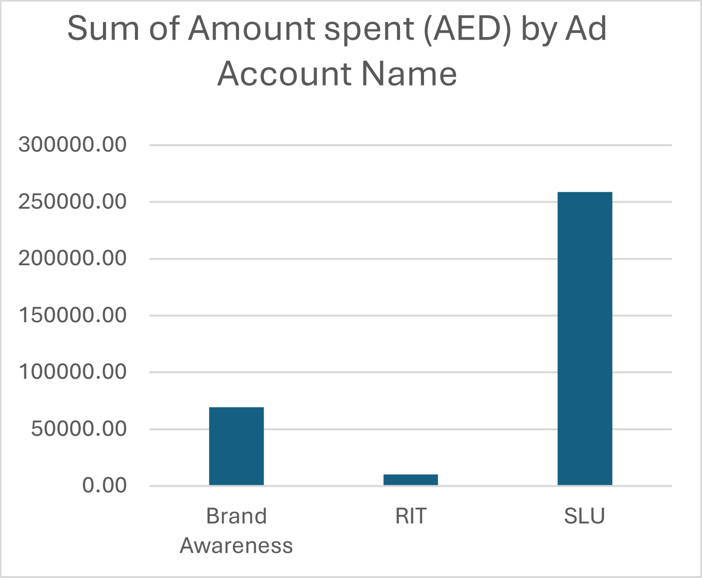
Identification of missing values, duplicates, and inconsistencies:

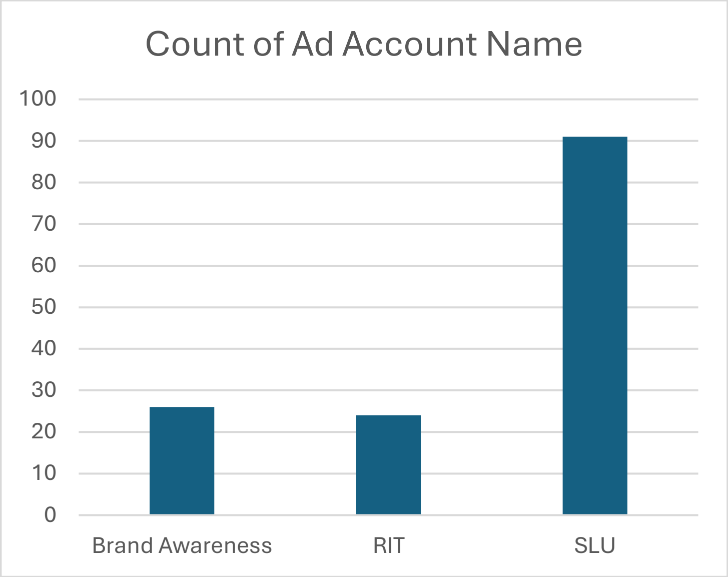
Missing Values: 08 (there are over all 8 missing values in the entire table(blanks))

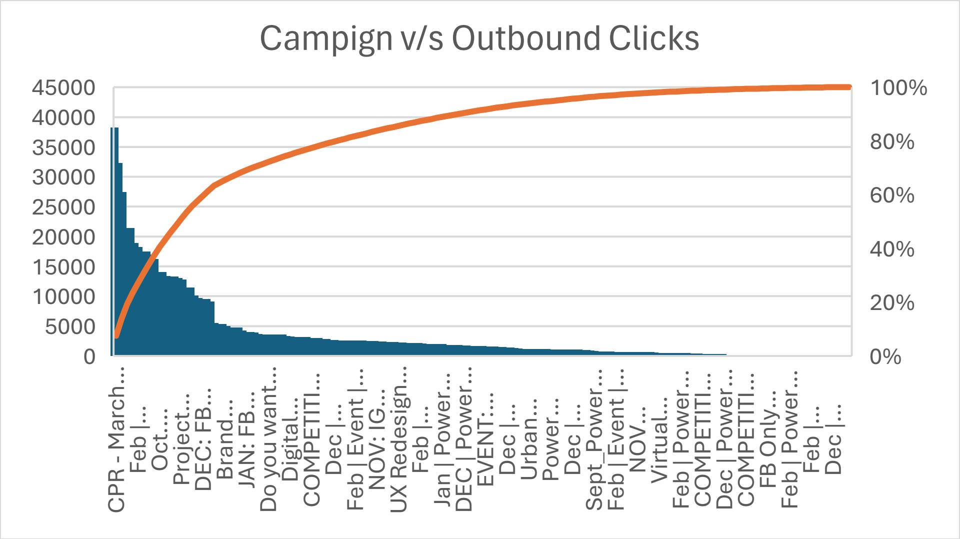
Duplicates: no duplicates

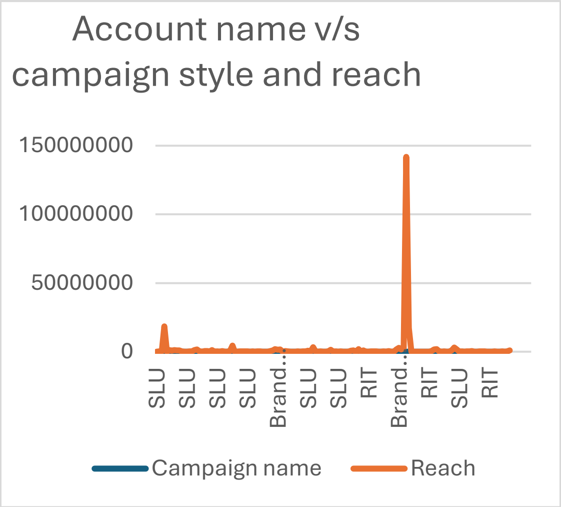
Inconsistencies:

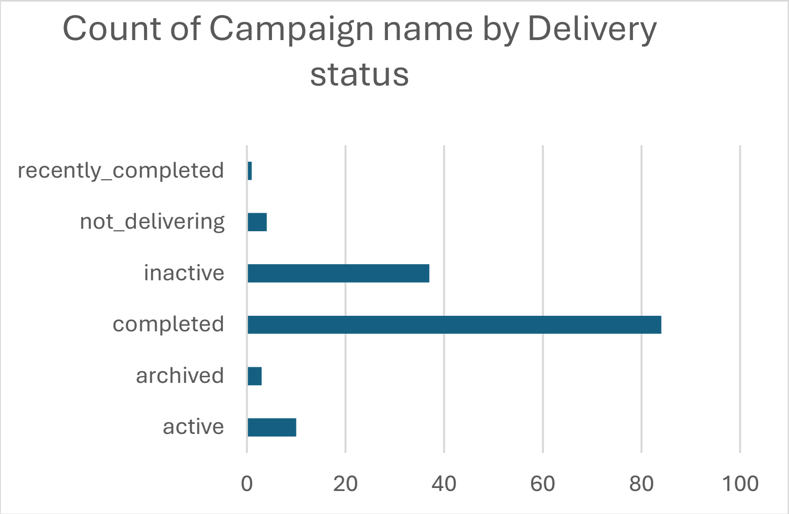
a) There are consistent in the date, like different format than expected, have anomalies, outliers in cost per result, amount, results.











Key Findings

Distribution:

- Most metrics (Reach, Results, Amount Spent, etc.) are stored as **TEXT** in the raw file, preventing direct numeric analysis until casting/cleaning.
- Initial inspection suggests multiple repeated Ad Account Name values; Campaign name varies but needs normalization (case, extra spaces).
- Outbound clicks and CPC have very limited distinct values in the raw stats output, indicating possible data parsing/formatting issues.

By Campaign:

- No declared primary key — multiple campaigns can share the same reporting date and ad account.
- Campaign naming inconsistencies (e.g., spacing, casing) could cause false duplicates when grouping.

Temporal:

- dates column present but needs validation — ensure it is a true DATE type and not a text string.
- Cannot yet compute start–end ranges or trends until data cleaning.

Quality:

- Many columns show **0 non-null counts** in your posted stats output — likely due to import or quoting issues when using \copy from CSV.
- Potential duplicates across Ad Account Name + Campaign name + dates need checking once cleaned.
- Numeric fields (Reach, Results, Amount Spent, etc.) are text in raw import — requires conversion before meaningful analysis.

Next Steps (for Week 2 Transformation)

1. **Data Type Fixes**
   - Cast Reach, Outbound clicks, Results, Cost per result, Amount spent (AED), CPC (cost per link click) to NUMERIC.
   - Cast dates to DATE type (MM/DD/YYYY parsing).
   - Ensure Ad Account Name, Campaign name, Delivery status, etc. remain as TEXT.

2. **Cleaning & Standardization**
    - Trim whitespace, normalize case in Campaign name and Ad Account Name.
    - Replace blank strings ('') with NULL where applicable.
    - Remove any non-numeric characters from numeric columns before casting.
3. **Key & Constraints**
    - Create a **composite primary key** on (ad_account_name, campaign_name, dates) if business rules support it.
    - Consider unique index on (campaign_name, dates) if each campaign has one record per date.
4. **Feature Engineering**
    - Add spend_per_click = amount_spent_aed / outbound_clicks.
    - Add result_rate = results / reach.
    - Derive campaign duration once grouped by campaign.
5. **Validation Rules**
    - Flag negative or zero metrics where not expected.
    - Flag future dates or unrealistic spend/reach values.
6. **Analytics Prep**
    - Create an indexed cleaned table (marketing_campaign_clean).
    - Build a view with essential metrics, standardized names, and derived features for direct use in Power BI/Tableau.

1. Dataset Overview

Source: Provided CSV file (Learner_Raw(in).csv)

Key Attributes:

- learner_id – Unique identifier for each learner
- country – Country of the learner
- degree – Current educational qualification
- institution – Name of educational institution
- major – Field of study

2. Summary Statistics (Categorical Overview)

Column	Non-Null Count	Unique Values	Most Frequent	Frequency
learner_id	129259	129259	Learner#00004f18-8b86-4fe4-ad7e-6c8d988f5335	1
country	126984	190	India	33868
degree	76566	7	Graduate Student	31806
institution	76358	34564	Saint Louis University	2163
major	76562	4502	Computer Science	4704

3. Missing Values & Duplicates

- learner_id: 0 (0.0%)
- country: 2275 (1.76%)
- degree: 52693 (40.77%)
- institution: 52901 (40.93%)
- major: 52697 (40.77%)

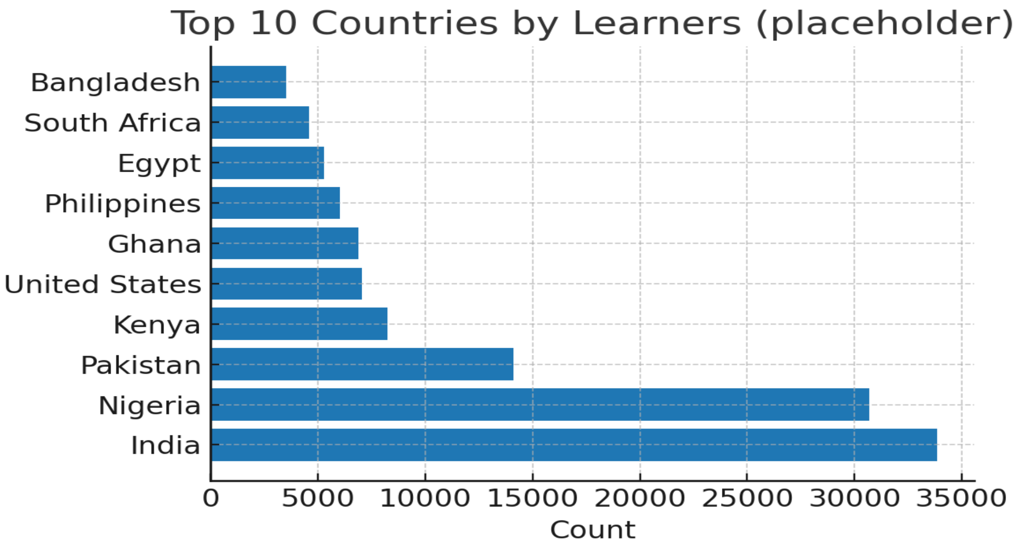
Duplicate Rows: 0

4. Inconsistencies Identified

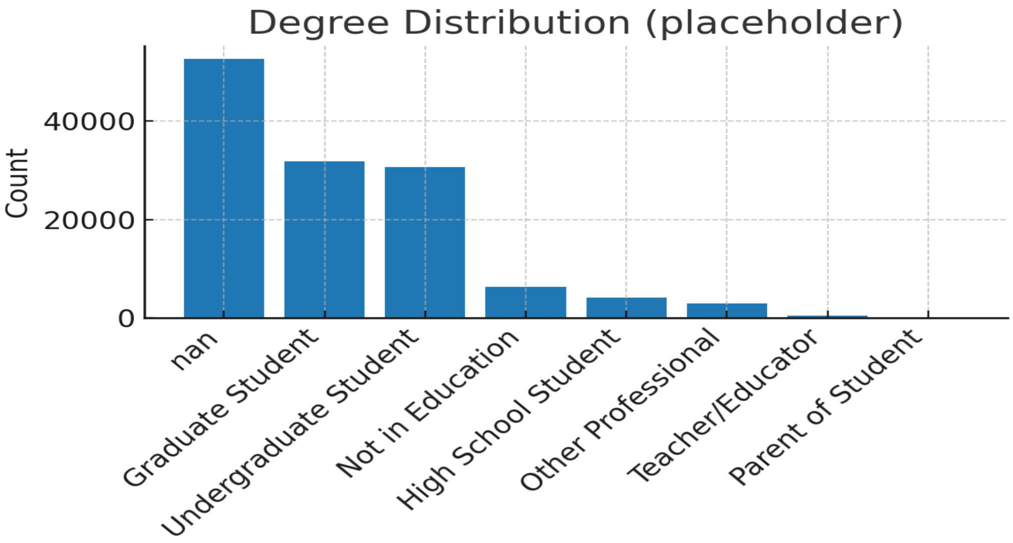
- Case variations in institution names (e.g., Saint Louis University vs saint louis university)
- Possible typos in institution names
- Empty strings in categorical columns

5. Visualizations (placeholders)

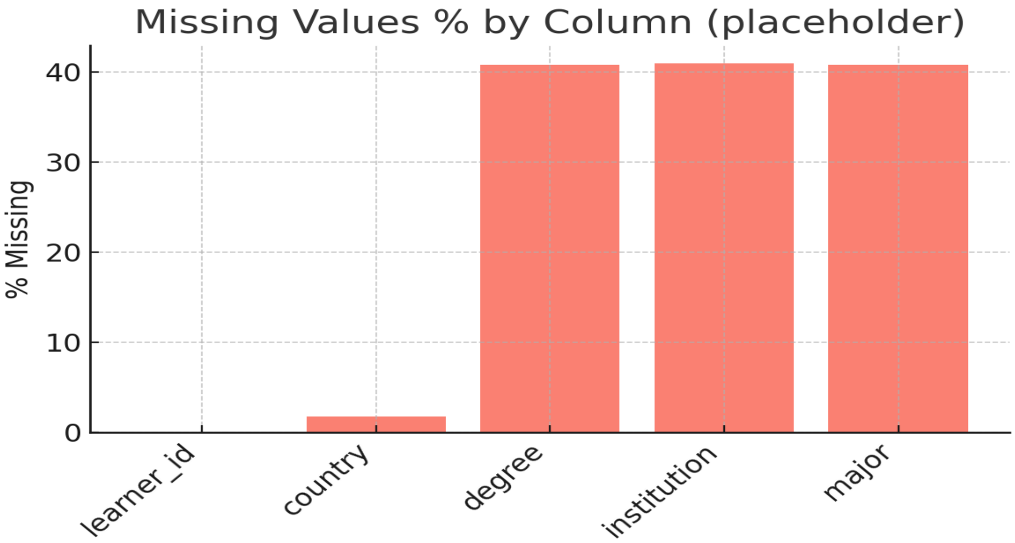
5.1 Country Distribution (Top 10)



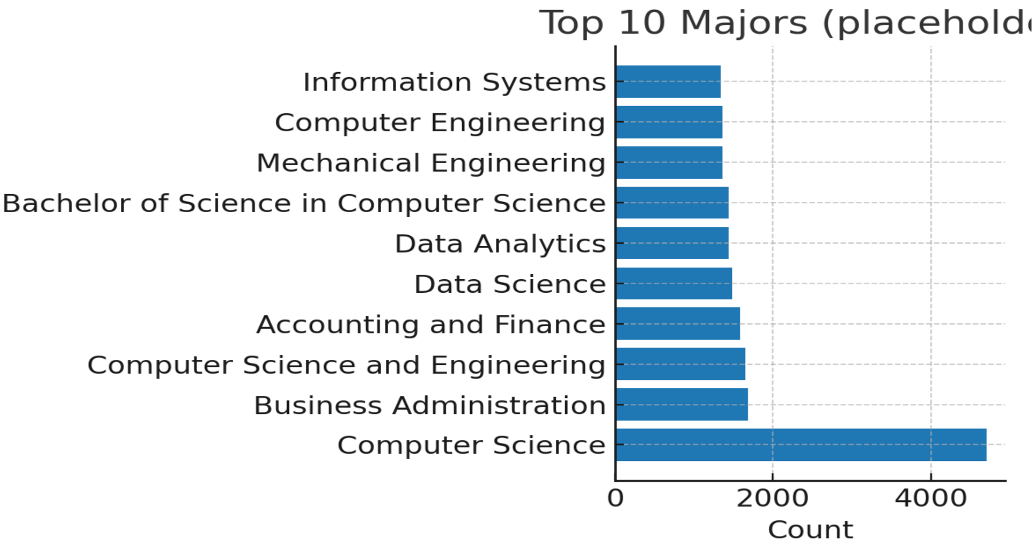
5.2 Degree Distribution



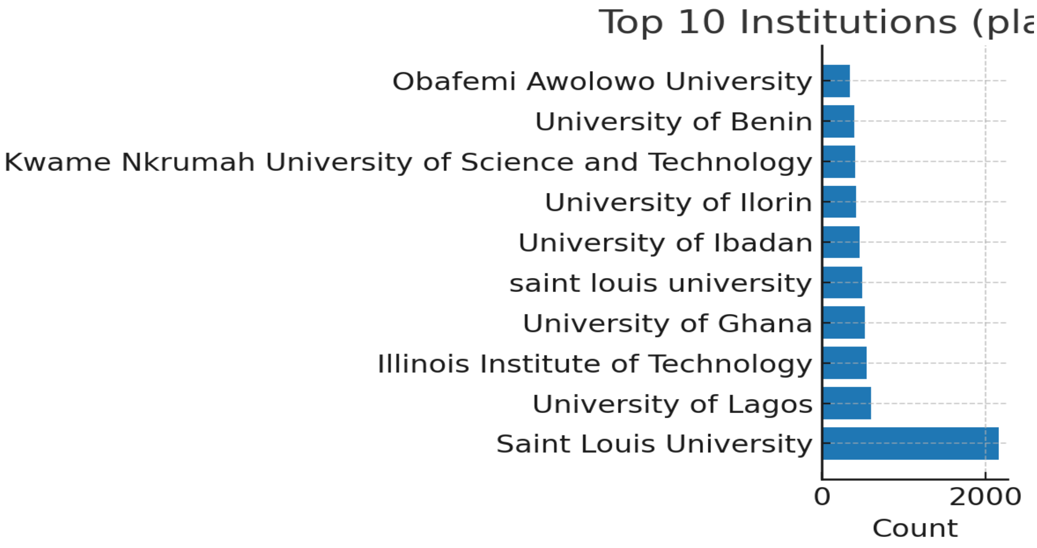
5.3 Missing Values (%)



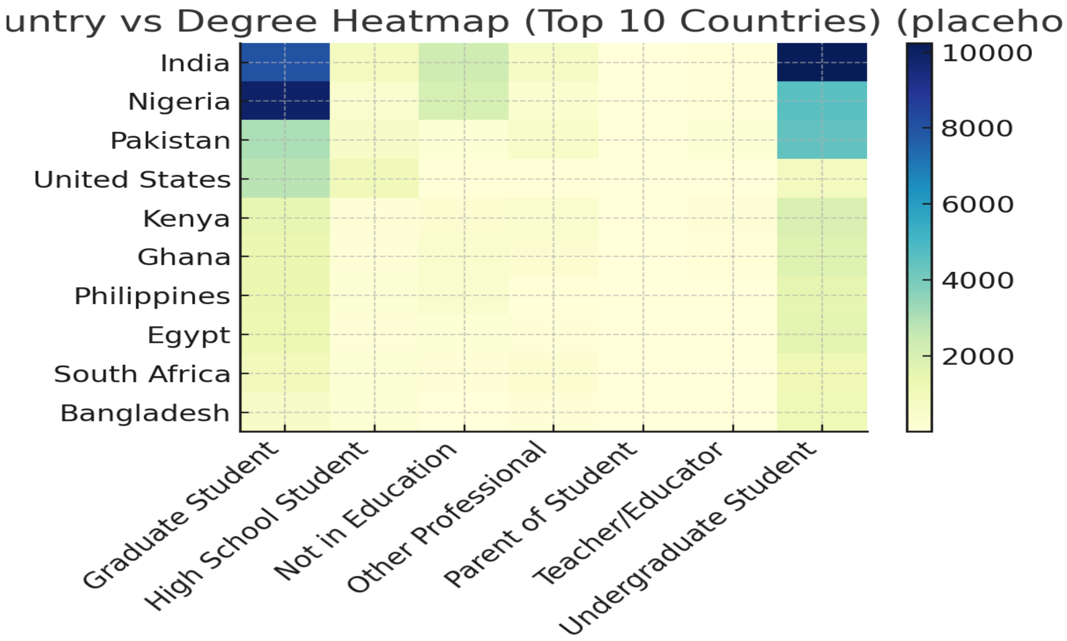
5.4 Major Distribution (Top 10)



5.5 Institution Frequency (Top 10)



5.6 Country vs Degree Co-occurrence Heatmap (Top 10 Countries)



6. Correlation Analysis

All variables are categorical; use co-occurrence heatmaps (e.g., Country vs Degree) to explore relationships.

7. Key Findings

- ~41% missing in 'degree', 'institution', and 'major'.
- 'country' mostly complete with top countries including India, Nigeria, and Pakistan.
- Graduate & Undergraduate students make up majority of the dataset.
- High uniqueness in institution names: standardization recommended.

8. Next Steps

- Convert empty strings to NULLs and standardize text (lowercase/trim).
- Decide on strategy for missing values (drop/impute/keep).
- Group rare categories under 'Other' where appropriate.
- Prepare cleaned dataset for Week 2 transformations and dashboards.

4. Learner_Opportunity Dataset

Column	Data Type	Non-Null Count
Enrollment_id	Object	113417
opportunity_id	Object	113603
assigned_cohort	Object	100284
apply_date	Datetime	11,3414
status	Int	11,3416

Summarize the Data

Since all fields are categorical,

we’ll analyze:

- Unique value counts

- Most frequent entries

Top 5 Application Status Code

Status Code	Count
1070	76109
1030	12236
1055	11471
1120	9048
1110	1514

Top 5 Assigned Cohorts

Cohort	Count
BAM6HBR	1805
BSEV9QO	1733
BGRQZ2N	1719
BP9ZV19	1611
BWAG78I	1564

Identify Missing and Duplicate Values

We'll now compute:

- Missing values per column

- Duplicate records Missing and Duplicate Data Summary

Column	Missing Count
assigned cohort	13318
enrollment_id	186
learner_id	188
Duplicate	0

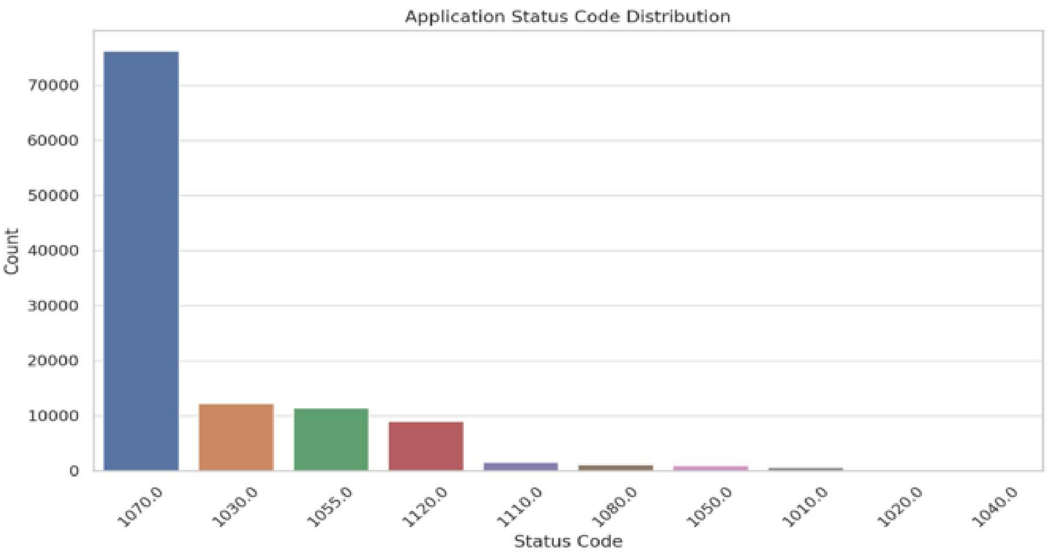
Spot Outliers and Anomalies

- Detected early timestamps in apply_date.
- Status values consistent with expected ranges.

Next, let's visualize the data distribution and missing values

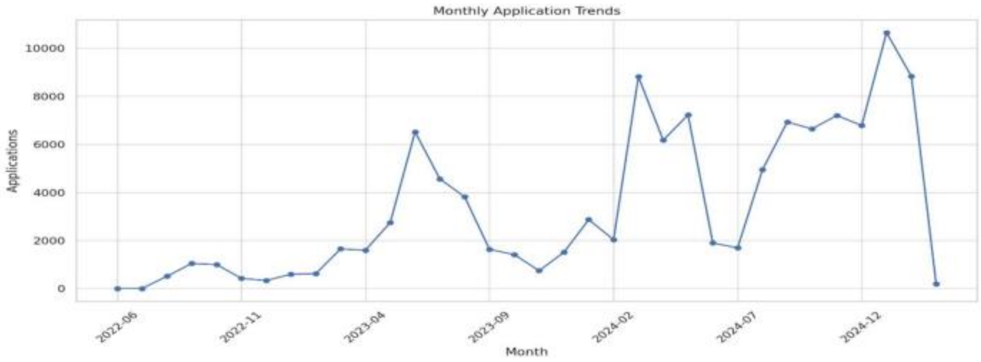
Application Status Code Distribution

- Shows how many records fall under each status code.
- Status 1070 dominates with over 76k applications, indicating the most common learner state



Monthly Application Trends

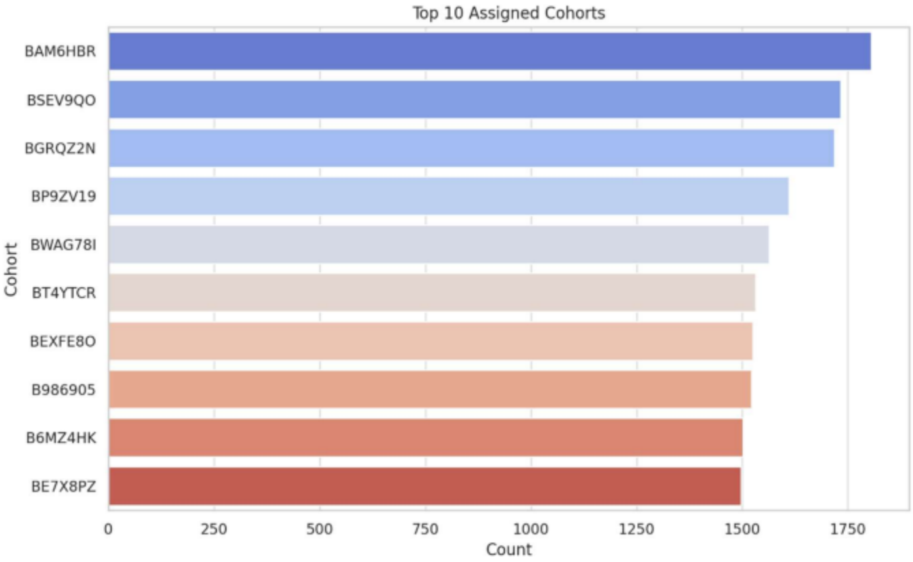
- Line chart showing application counts by month.
- Major spike observed around May 2024, showing seasonal activity or campaign influence.



Top 10 Assigned Cohorts

- Bar chart showing most frequent cohort assignments.

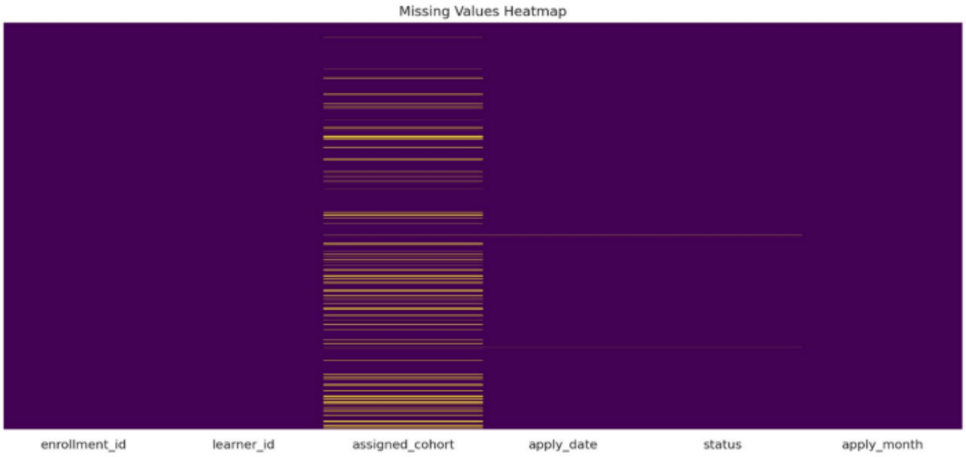
- Indicates strong clustering with top cohorts like BAM6HBR and BSEV9QO



Missing Values Heatmap

- Visualizes null values across all columns.

- Clearly highlights gaps in assigned_cohort, enrollment_id, and learner_id



Visual Insights

- Status code 1070 dominates the dataset, indicating the most frequent learner outcome or enrollment state.

- Application activity peaked in May 2024, showing potential alignment with campaigns or seasonal trends.

- A small number of cohorts (e.g., BAM6HBR, BSEV9QO) account for a large portion of learner participation.

- The missing values heatmap shows notable gaps in assigned_cohort, enrollment_id, and learner_id.

2.Opportunity_Raw Data Set

Column	Data Type	Non-Null Count
opportunity_id	Object	187
opportunity name	Object	187
category	Object	187
opportunity code	Object	187
tracking questions	Object	187

Summarize the Data

Since all fields are categorical, we'll analyze:

- Unique value counts
- Most frequent entries

Top 6 Opportunity Categories

Category	Count
Internship	43
Event	41
Competition	41
Career	23
Course	18
Masterclass	11
Engagement	10

Identify Missing and Duplicate Values

We'll now compute:

- Missing values per column
- Duplicate records

Missing and Duplicate Data Summary

- Missing Values: None
- Duplicate Rows: None

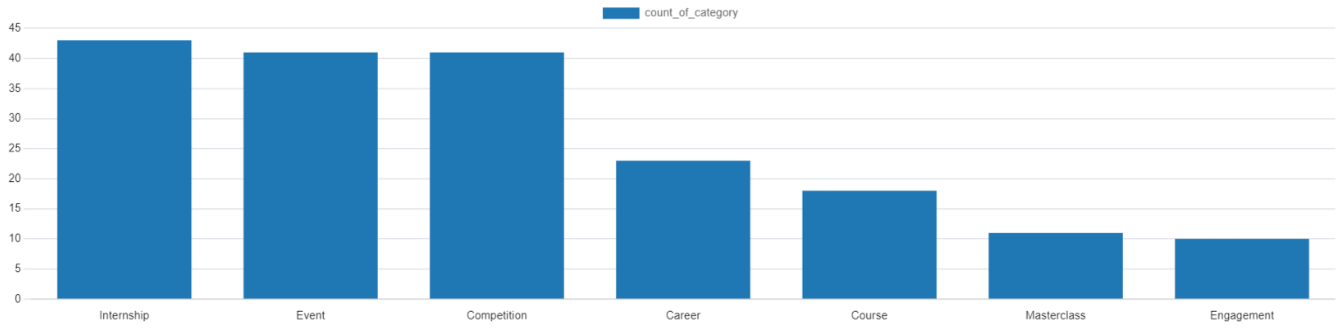
Spot Outliers and Anomalies

- No numeric columns to test for outliers.
- All categories are valid.

Next, let's visualize the data distribution and missing values.

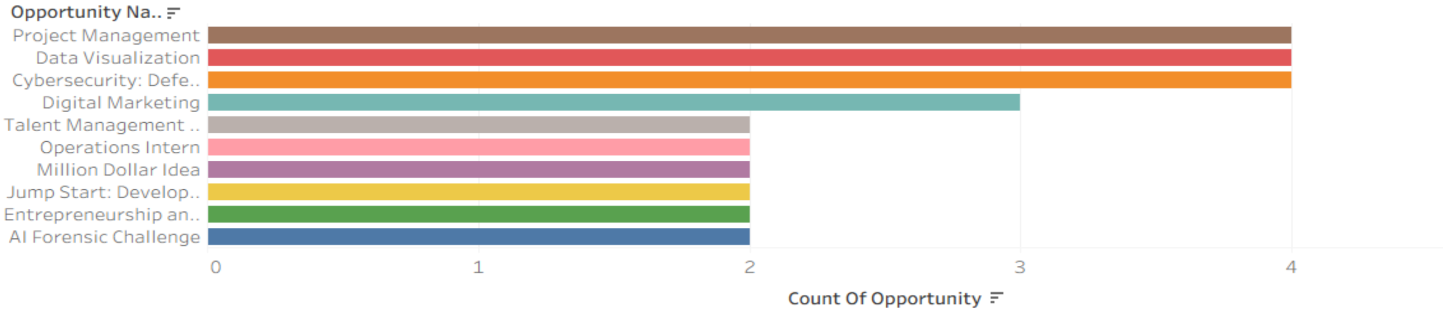
Opportunity Category Distribution

- Bar chart showing frequency of opportunity types.



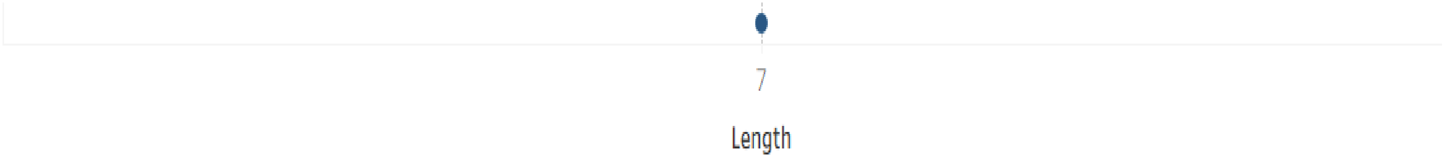
Top 10 Opportunity Names

- Highlights the most common or repeated opportunity titles.
- Useful for identifying popular or duplicated program offerings.



Opportunity Code Length

- Box plot showing variability in the length of opportunity_code.
- Confirms codes are consistently formatted with no major outliers.

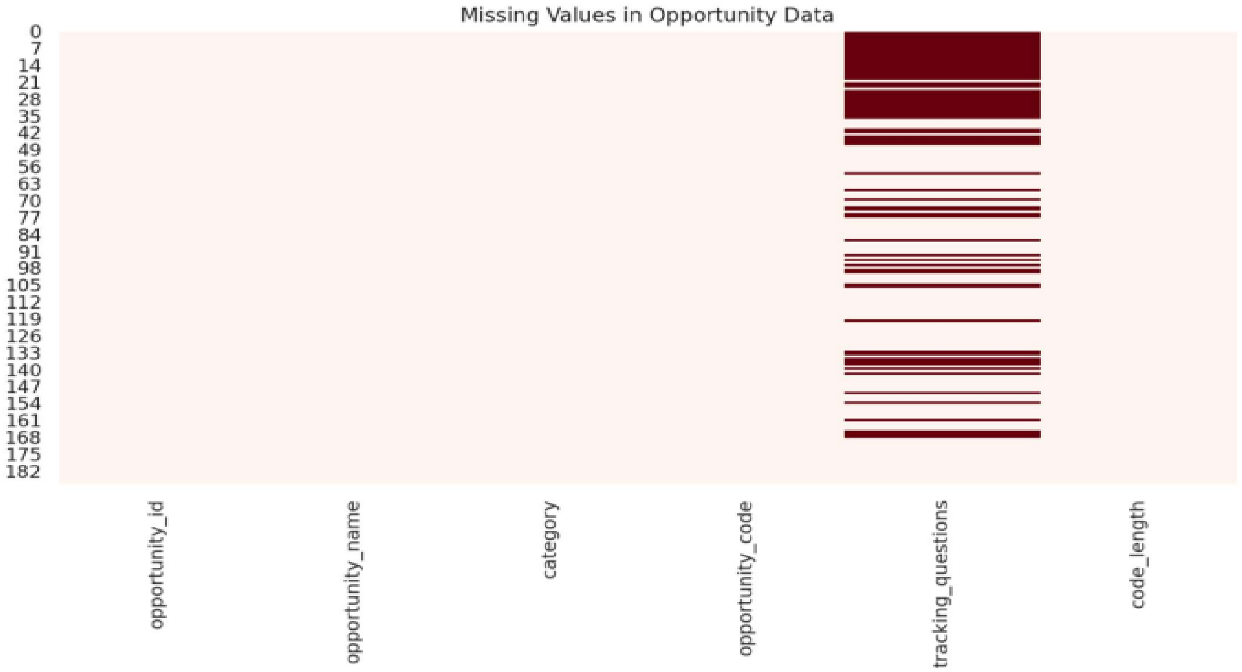


.

Missing Values Heatmap

- No missing values; chart appears clean and complete.

.● Validates high data quality for this dataset.



Visual Insights – Opportunity Dataset

- Most opportunities are categorized under Internship, Event, or Competition, reflecting high engagement programs.
- Repeated opportunity names suggest recurring programs or potential naming inconsistencies.
- Opportunity codes are consistently formatted, as shown in the code length bubble picture.
- Heatmap shows no missing values, indicating a clean and reliable dataset for immediate use.