Data Visualization Associate Early Internship

Week 1: Exploratory Data Analysis (EDA) Report

by

0303 DVA TEAM 22B

- **Submitted by** **Areeba Fatima (areebafatima721@gmail.com)**
- **Team Members:**

Sakshi Gollar (sakshigollar31@gmail.com)

Wamiq Ejaz (ejazwamiq@gmail.com)

Shivani Galande (shivanigalande2512@gmail.com)

Areeba Fatima (areebafatima721@gmail.com)

Varun D (varundevaraj1188@gmail.com)

Date: 10/03/25

1. User Data (user_data.csv)

By **Varun D (varundevaraj1188@gmail.com)**

1) Overview of the dataset structure, sources, and key attributes:

Dataset Structure and Sources

- **user_data.csv**

Datasets contain **129,259 rows and 5 columns**, primarily related to learners' education details.

- learner_id: Unique identifier for each learner.
- Country: The country of the learner.
- Degree: The degree obtained.
- Institution: The institution awarding the degree.
- Major : Field of study

2) Summary statistics of key variables:

Column	Unique Values	Most Common Value	Missing Values
learner_id	129,259	Unique for all	0
Country	190	India (33,868)	2,275 (Dataset 2)
Degree	8	"Null" (52,693)	0
Institution	27,036	"Null" (52,693)	157
Major	4,250	"Null" (52,697)	0

Observations:

- Many missing or "Null" values in Degree, Institution, and Major.
- learner_id is unique with no duplicates.

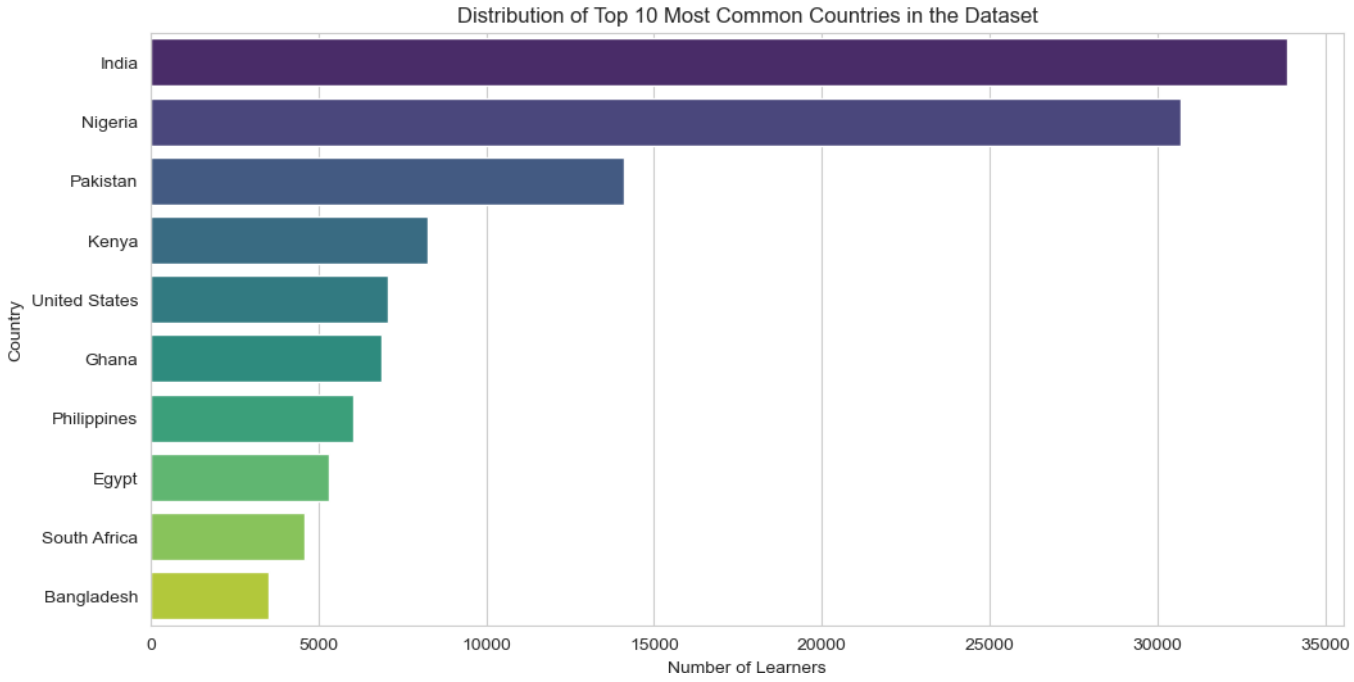
3) **Identification of missing values, duplicates, and inconsistencies:**

- Institution has **157 missing values**.
- Country has **2,275 missing values** (Dataset 2 only).
- "Null" values in Degree, Institution, and Major may need standardization.
- No duplicate learner_id values found.

4) **Data visualizations:**

(A) Distribution of Countries

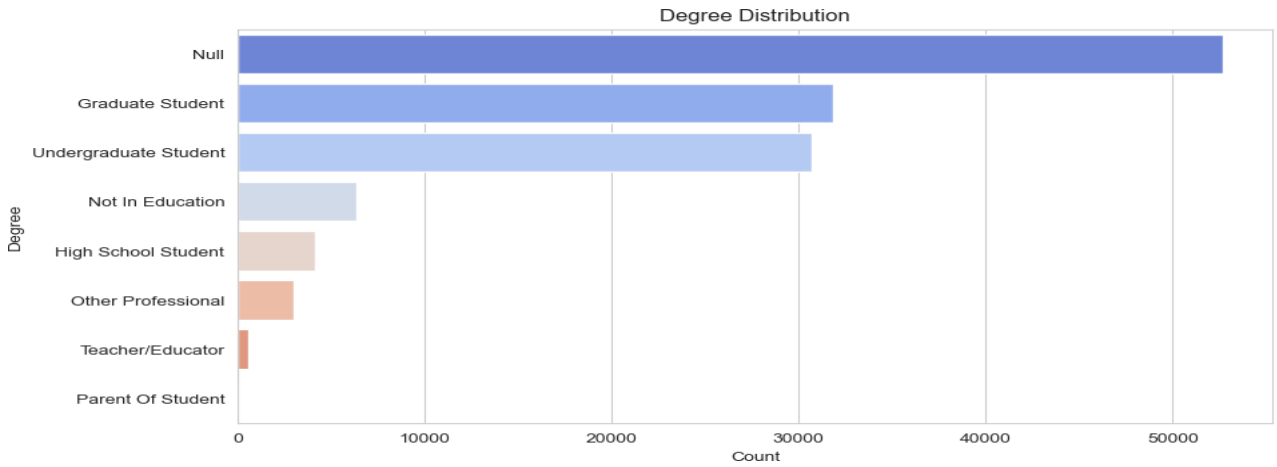
- India is the most represented country (26% of total data).
- A barplot would show the distribution of learners across different countries.



(B) Degree Distribution

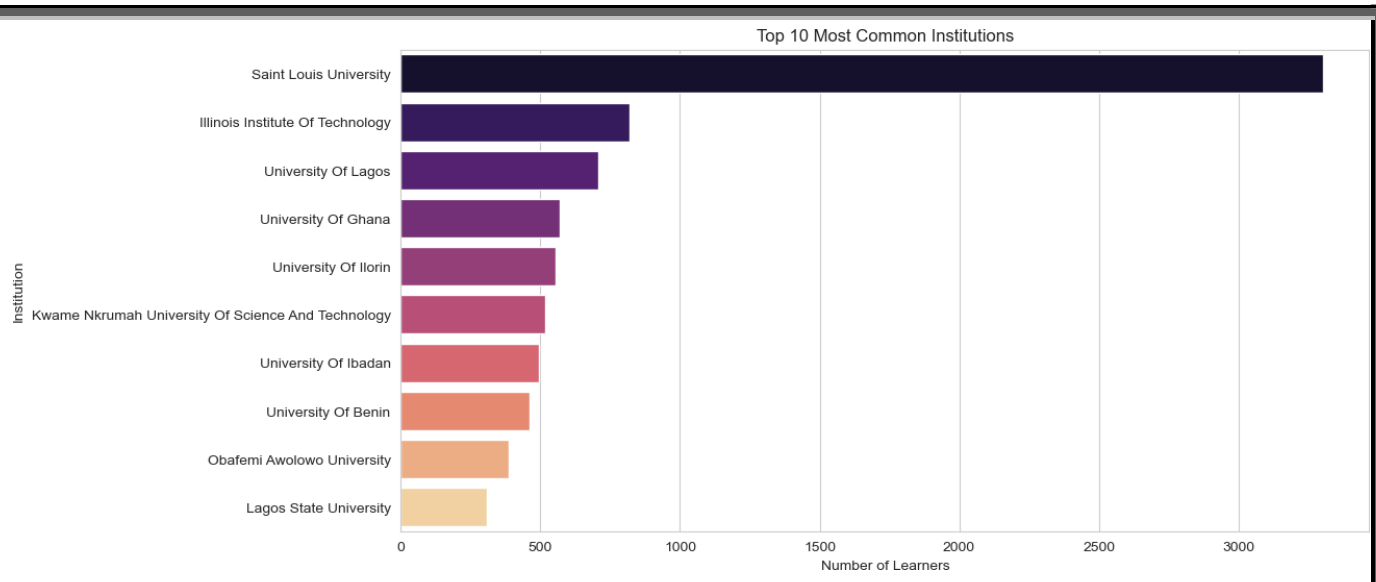
- Degrees have only **8 unique values**, with "Null" being the most common.

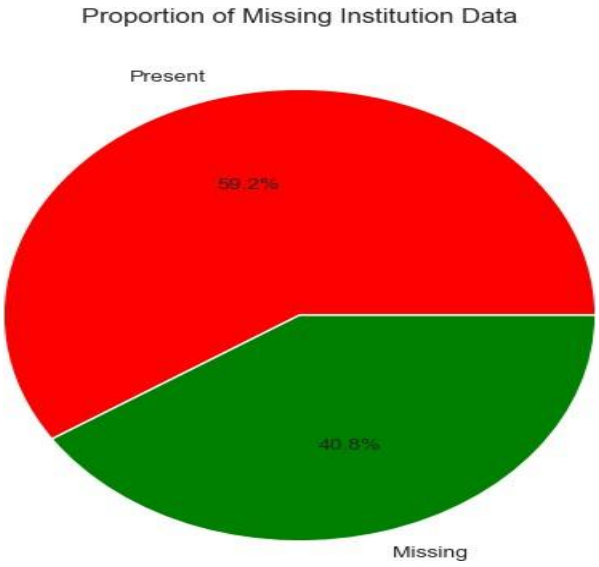
- A countplot could visualize the distribution.



(C) Institution Variability

- Over **27,000 unique institutions**.
- The presence of "Null" suggests inconsistencies in data entry.





5) Key findings and next steps for data cleaning and transformation:

Key Findings:

- High missing/null values in key fields.
- Country data mismatch between datasets.
- Large variety of institutions, requiring potential standardization.

Next Steps for Data Cleaning & Transformation:

- Replace "Null" with proper NULL values.
- Investigate and fill missing `Country` values where possible.

- Standardize column names (`Major` vs. `Degree.1`).
- Normalize institution names to avoid duplicates (e.g., "MIT" vs. "Massachusetts Institute of Technology").
- Perform deeper analysis post-cleaning

2. Opportunity Data (opp_data.csv)

By **Shivani Galande (shivanigalande2512@gmail.com)**

1) Overview of the dataset structure, sources, and key attributes:

- Dataset Name: Opportunity_Raw
- Data sources: CRM or database export and appears to be historical.
- Key Attributes

Column Name	Description	Data Type	Key Observations
opportunity_id	Unique identifier for each opportunity	Categorical (ID)	Likely a **Primary Key (PK)**
opportunity_name	Name of the opportunity	Categorical (Text)	Needs standardization (case/spacing issues)
category	Type of opportunity (e.g., Event, Internship)	Categorical	Some categories may need grouping
opportunity_code	Another unique identifier (might be alternative PK)	Categorical (ID)	Check uniqueness
tracking_questions	Might store JSON-like tracking data	jsonb	Contains **missing values (36.9%)**

2) Summary statistics of key variables:

- Number of rows: 187
- Number of columns: 5
-

category	count
Internship	43
Event	41
Competition	41

Career	23
Course	18
Masterclass	11
Engagement	10

- Opportunity Code length = 7
- Opportunity id length = 38

3) Identification of missing values, duplicates, and inconsistencies:

- Missing Values

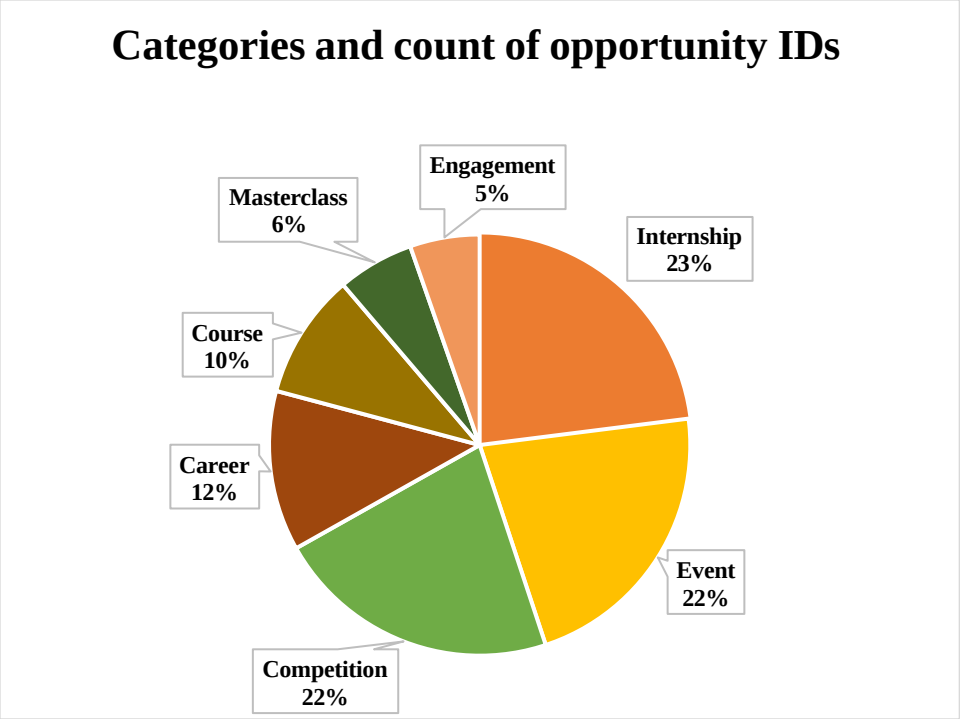
Column Name	Total Values	Non-null Values	Missing Values
opportunity_id	187	187	0
opportunity_name	187	187	0
category	187	187	0
opportunity_code	187	187	0
tracking_questions	187	118	69

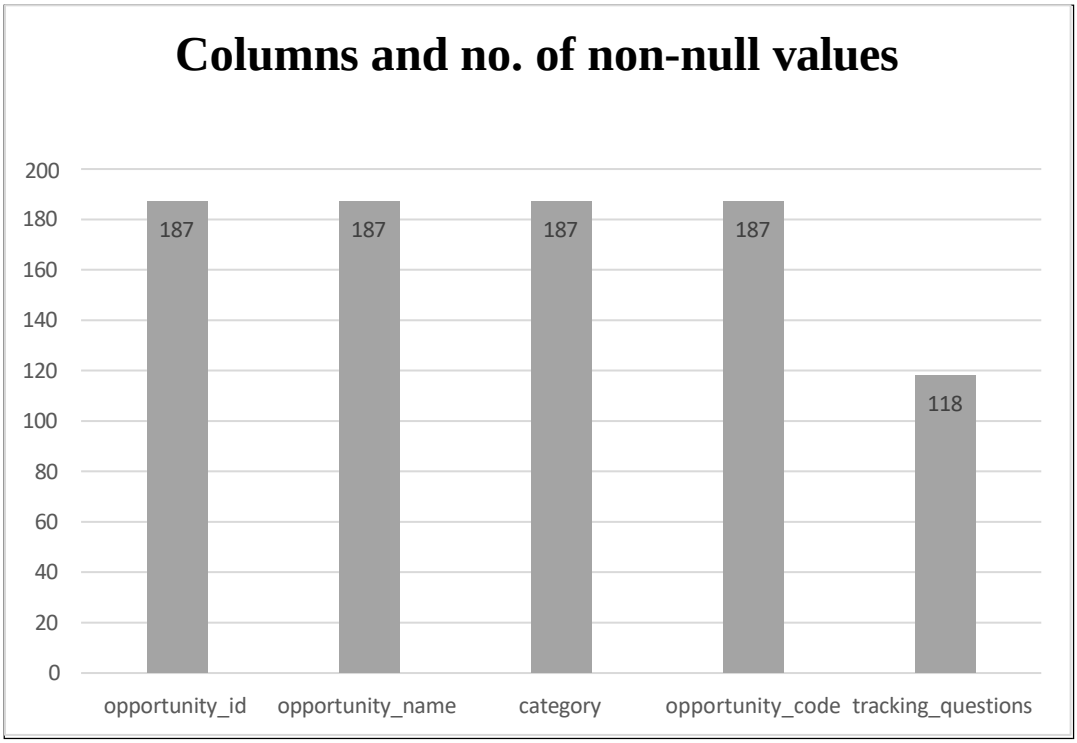
- Duplicates:

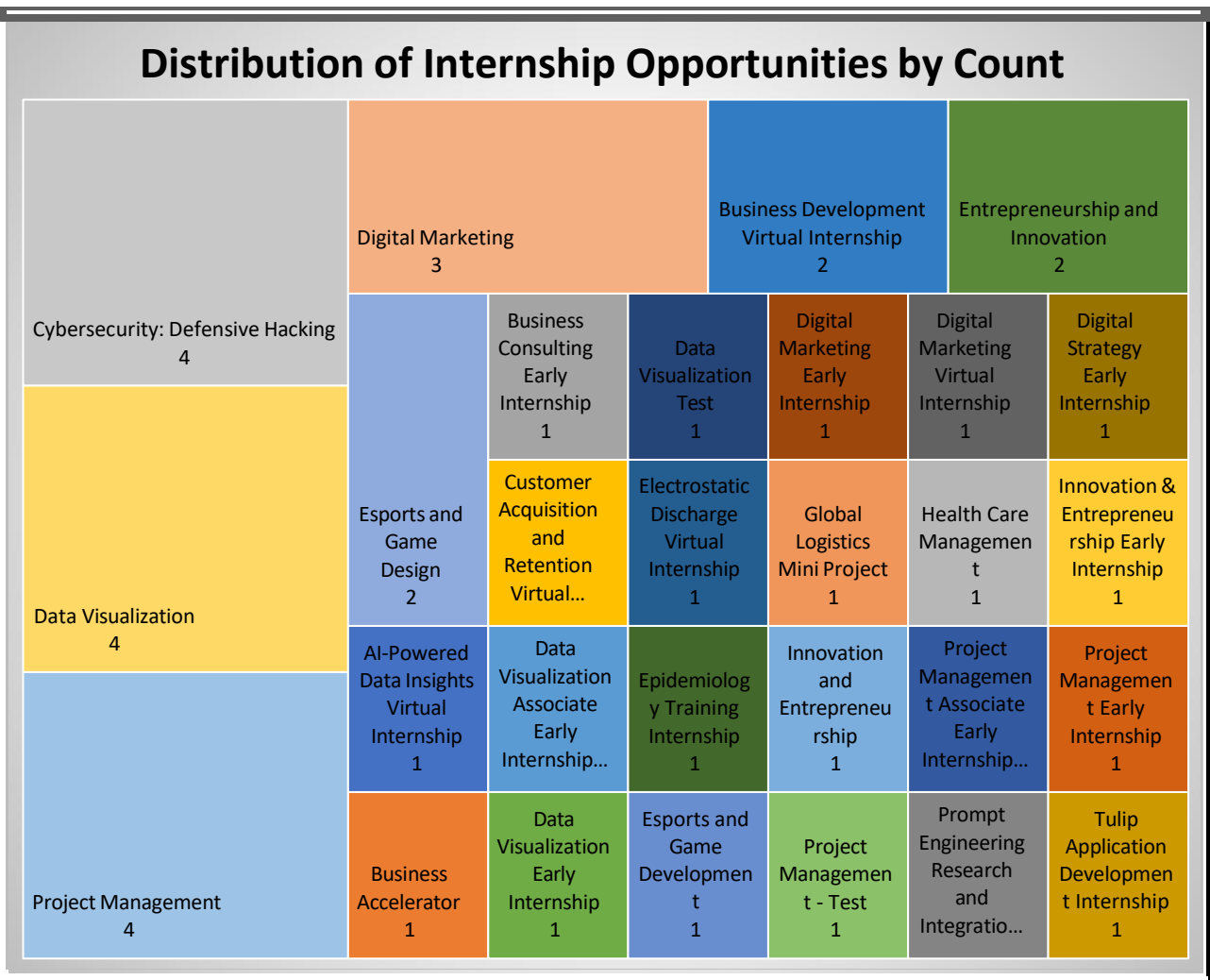
Column Name	Uniques Observations
opportunity_id	187
opportunity_name	170
category	7
opportunity_code	187

- Inconsistencies and Outliers: In columns,
  - opportunity_code: None
  - category: None
  - tracking_questions: Some values contained NULL as a string instead of actual NULL values, which were replaced by “Not Available”

4) Data visualizations:







5) Key findings and next steps for data cleaning and transformation:

- Findings:
  - No duplicate values were found.
  - Tracking questions column has a significant number of missing values.
  - Data consistency is maintained in opportunity_id, opportunity_name, and category.
  - No outliers or unexpected characters in opportunity_code.
- Next Steps:
  - Handle missing values in tracking_questions (either remove rows or replace NULL values based on business logic).
  - Ensure all categories are correctly classified to avoid misinterpretations.
  - Perform JSON Flattening for extracting insights from tracking_questions.

3. Cohort Data (cohort_data.csv)

By **Wamiq Ejaz (ejazwamiq@gmail.com)**

1) Overview of the dataset structure, sources, and key attributes:

- Dataset Name: Cohort Data

The dataset used in this analysis, **Cohort Data (cohort_data.csv)**, tracks cohort-based learning programs, including:

- Cohort sizes
- Timelines
- Linked opportunities
- Participation and completion analysis
- Understanding this dataset allows us to uncover trends in learning engagement and outcomes.

- *Data Source*

- Origin: Internal data source
- Frequency of Updates: Historical data
- Potential Biases: Possible inconsistencies in cohort size reporting

2) Summary statistics of key variables:

Below are key summary statistics of the dataset:

ATTRIBUTE	DATA TYPE	MINIMUM	MAXIMUM	MEAN	STD DEVIATION
COHORT SIZE	Integer	3	100,000	5741	20994
START DATE	Date	09/06/2022	07/08/2025	-	-
END DATE	Date	10/06/2022	06/03/2026	-	-

3) Identification of missing values, duplicates, and inconsistencies:

- Missing Values

  - No missing values were detected in the dataset.

- Duplicates

  - No duplicate **records** were found.

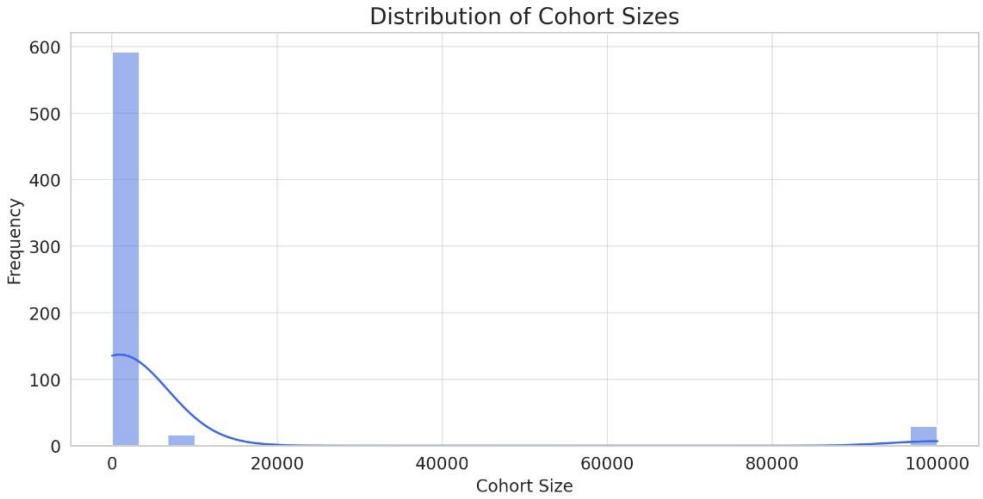
- Key findings:

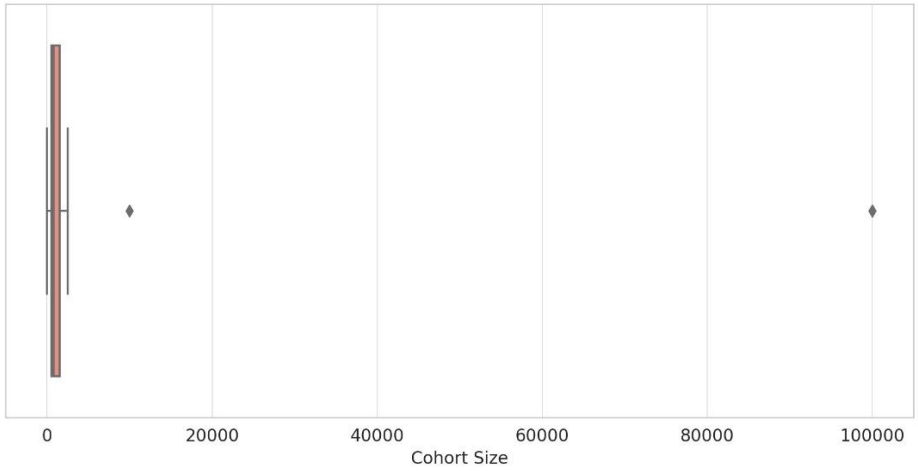
  - Cohort size anomalies: Several cohorts have exceptionally large sizes (e.g., 100,000 -30 entries), requiring verification.
  - High standard deviation: The cohort size has significant variation, impacting data consistency.

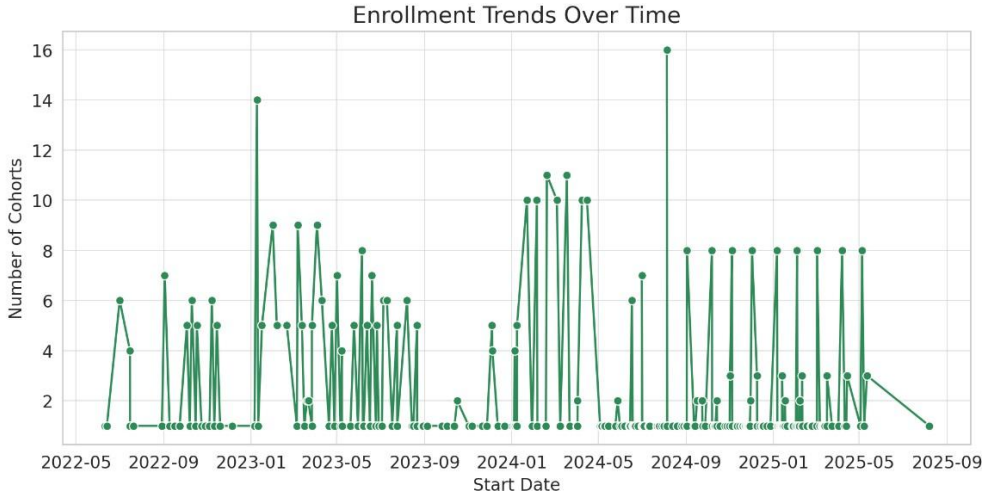
4) **Data visualizations:**

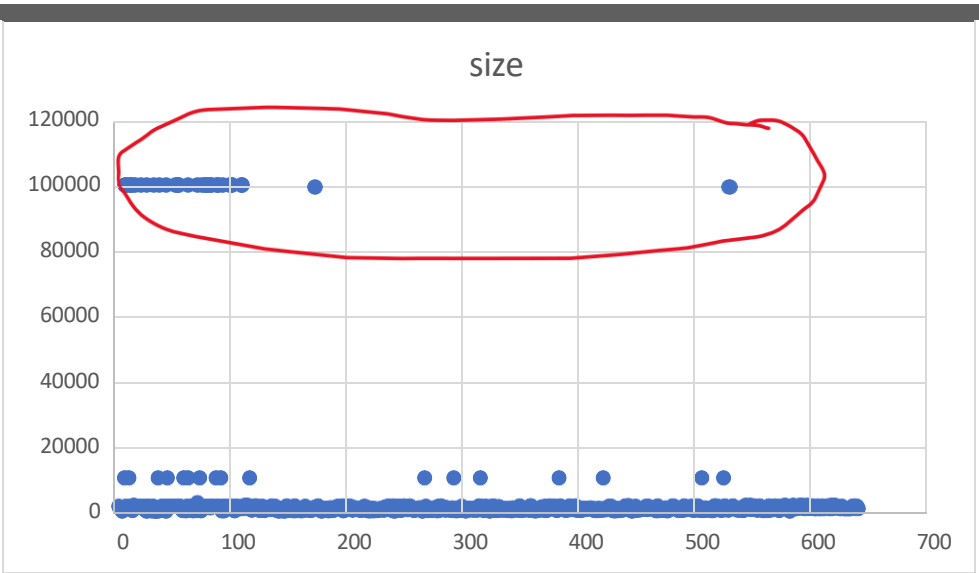
Visualizations used to understand the dataset:

- Histogram of Cohort Sizes: Shows extreme variability with some very large cohorts.
- Box Plot of Cohort Sizes: Highlights significant outliers.
- Line Chart of Enrolment Trends Over Time: Patterns in cohort start dates.
- Scatter Graph of Cohort Sizes highlights the outliers.


Distribution of Cohort Sizes
Frequency
Cohort Size
0
100
200
300
400
500
600
0
20000
40000
60000
80000
100000


Box Plot of Cohort Sizes
Cohort Size
0
20000
40000
60000
80000
100000


Enrollment Trends Over Time
Number of Cohorts
Start Date
2
4
6
8
10
12
14
16
2022-05
2022-09
2023-01
2023-05
2023-09
2024-01
2024-05
2024-09
2025-01
2025-05
2025-09



5) **Key findings and next steps for data cleaning and transformation:**

- Key Findings

- Cohort sizes show **significant variation**, with some exceptionally large cohorts.
- **No missing or duplicate** values were detected.
- The dataset is ready for further cleaning and transformation before visualization.

- Next Steps

- Verify and **handle outliers** in cohort sizes.
- Proceed with **data cleaning** to ensure accurate analysis.
- Prepare the cleaned dataset for transformation and visualization

4. Marketing Data (marketing_data.csv)

By **Sakshi Gollar (sakshigollar31@gmail.com)**

1) Overview of the dataset structure, sources, and key attributes:

The dataset analyzed contains marketing campaign performance metrics, including engagement, reach, and costs. It consists of **58 records** and **13 columns**, with a mix of categorical and numerical data.

2) Summary statistics of key variables:

Summary Statistics **Metric**	**Min**	**Max**	**Mean**	**Std Dev**
Count of Results	1	141	11.00	22.78
Sum of Reach	26,184	240,102,000	24,078,190	48,823,130
Sum of Outbound Clicks	456	511,965	43,070	85,685
Sum of Landing Page Views	2	293,842	23,580	51,557
Sum of Cost per Result	0.0085	587.11	42.69	103.26
Sum of CPC (Cost per Link Click)	0.27	147.93	11.15	24.01
Sum of Amount Spent (AED)	411.68	338,393.71	27,793.03	55,693.65

3) Identification of missing values, duplicates, and inconsistencies:

- Missing Values & Duplicates
  - "Delivery Status" (45 missing)

- "Ad Account Name" (36 missing)
- "Delivery Level" (47 missing)
- "Result Type" (28 missing)
- "Reporting Starts" (38 missing)
- "Reporting Ends" (48 missing)

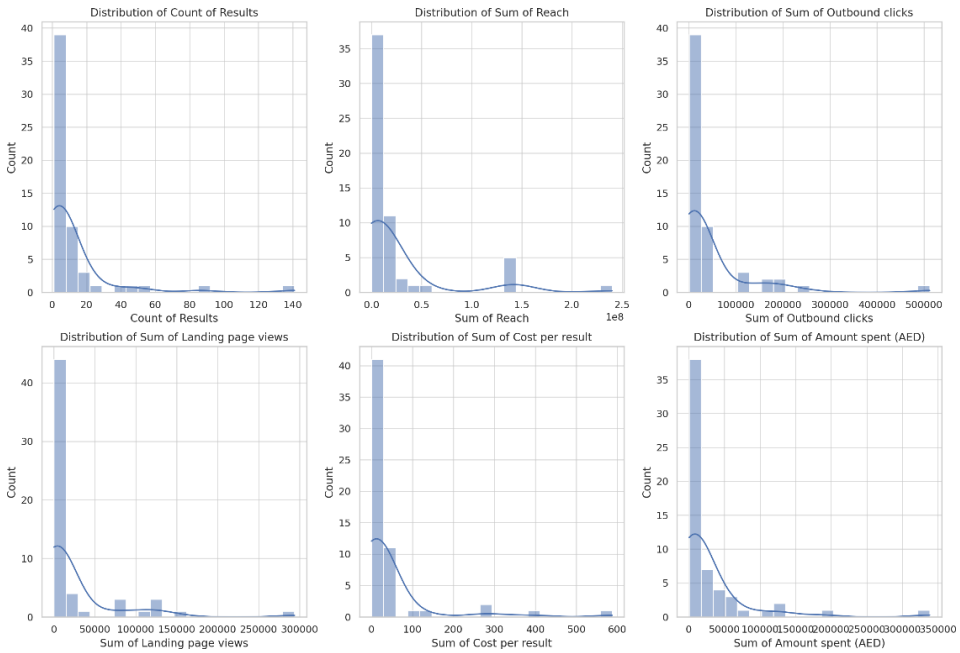
➢ Duplicates: None found.

➢ Outliers & Anomalies

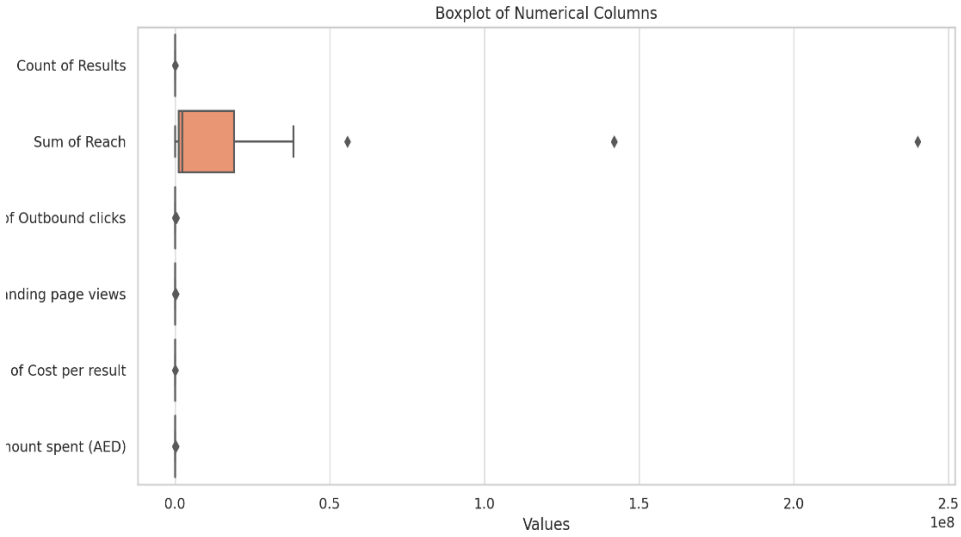
- Several columns exhibit high variance and extreme values, including "Sum of Reach," "Sum of Amount Spent (AED)," and "Sum of Cost per Result."
- Some campaigns recorded very high CPC (Cost per Click), indicating inefficiencies.
- Significant outliers in "Sum of Outbound Clicks" and "Sum of Landing Page Views" suggest inconsistencies in campaign performance.

4) Data visualizations:

- Histogram: Reveal skewed distributions in most numerical fields.



- Boxplots: Highlight extreme outliers, especially in cost-related metrics.



5) Key findings and next steps for data cleaning and transformation:

- Key Findings
  - Campaign performance varies widely, with some campaigns incurring **exceptionally high costs per result**.
  - There are **missing values** in key categorical columns, which may impact segmentation analysis.
  - The presence of **outliers** suggests the need for data cleaning or further investigation.
- Next Steps
  - Handle missing categorical data by **imputing or removing irrelevant columns**.
  - Investigate and potentially remove extreme outliers to ensure analysis accuracy.
  - Normalize cost metrics to make cross-campaign comparisons more meaningful.
  - Perform deeper segmentation to understand performance drivers in high-cost campaigns.

5. Learner Data (learner_raw.csv)

By **Varun D (varundevaraj1188@gmail.com)**

1) Overview of the dataset structure, sources, and key attributes:

Dataset Structure and Sources

- **Learner_RawTrimed.csv**

Datasets contain **129,259 rows and 5 columns**, primarily related to learners' education details.

Key Attributes:

- learner_id: Unique identifier for each learner.
- Country: The country of the learner.
- Degree: The degree obtained.
- Institution: The institution awarding the degree.
- Major Degree.1 : Field of study.

2) Summary statistics of key variables:D

Column	Unique Values	Most Common Value	Missing Values
learner_id	129,259	Unique for all	0
Country	190	India (33,868)	2,275
Degree	8	"Null" (52,693)	0
Institution	27,036	"Null" (52,693)	157
Degree.1	4,250	"Null" (52,697)	0

Observations:

- Many missing or "Null" values in Degree, Institution, and Degree.1.
- learner_id is unique with no duplicates.

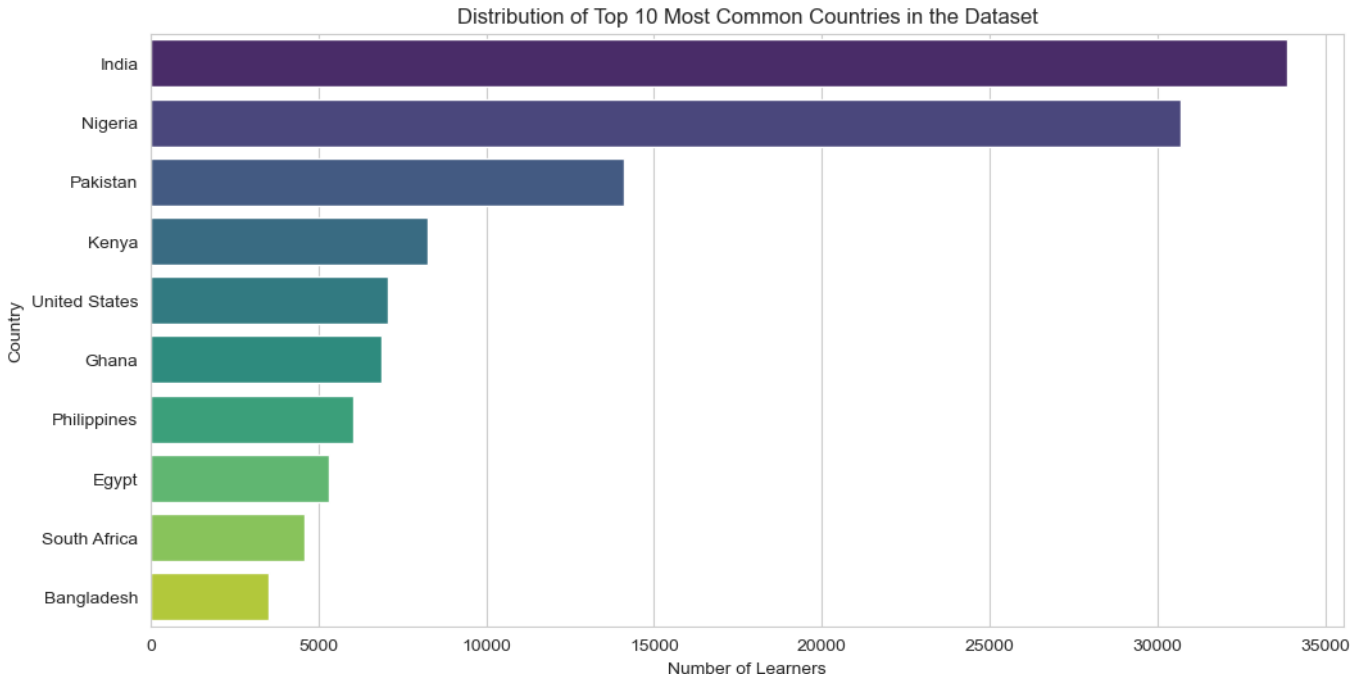
3) Identification of missing values, duplicates, and inconsistencies:

- Institution has **157 missing values**.
- Country has **2,275 missing values** (Dataset 2 only).
- "Null" values in Degree, Institution, and Degree.1 may need standardization.
- No duplicate learner_id values found.

4) Data visualizations:

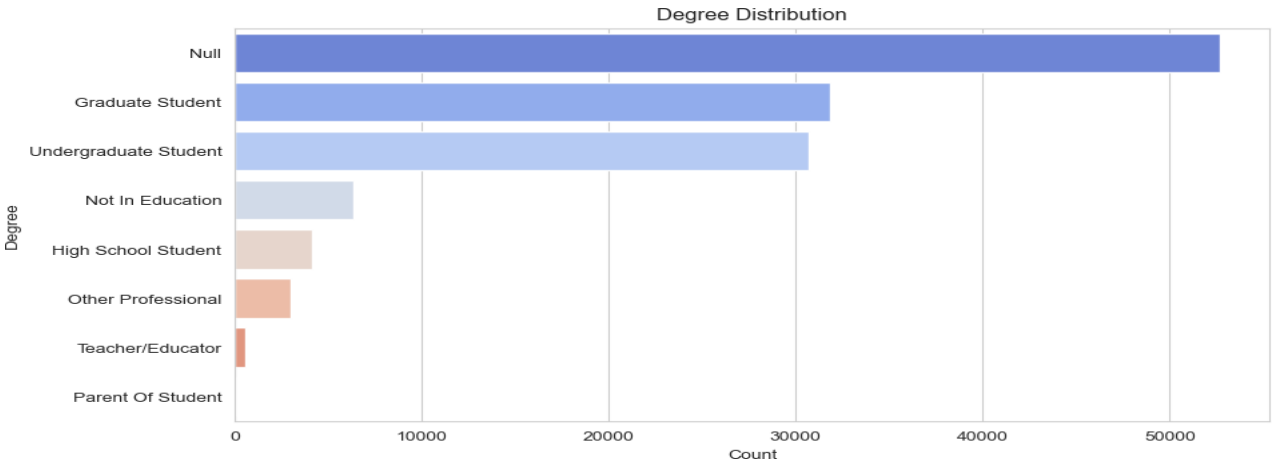
(A) Distribution of Countries

- **India** is the most represented country (26% of total data).
- A barplot would show the distribution of learners across different countries.



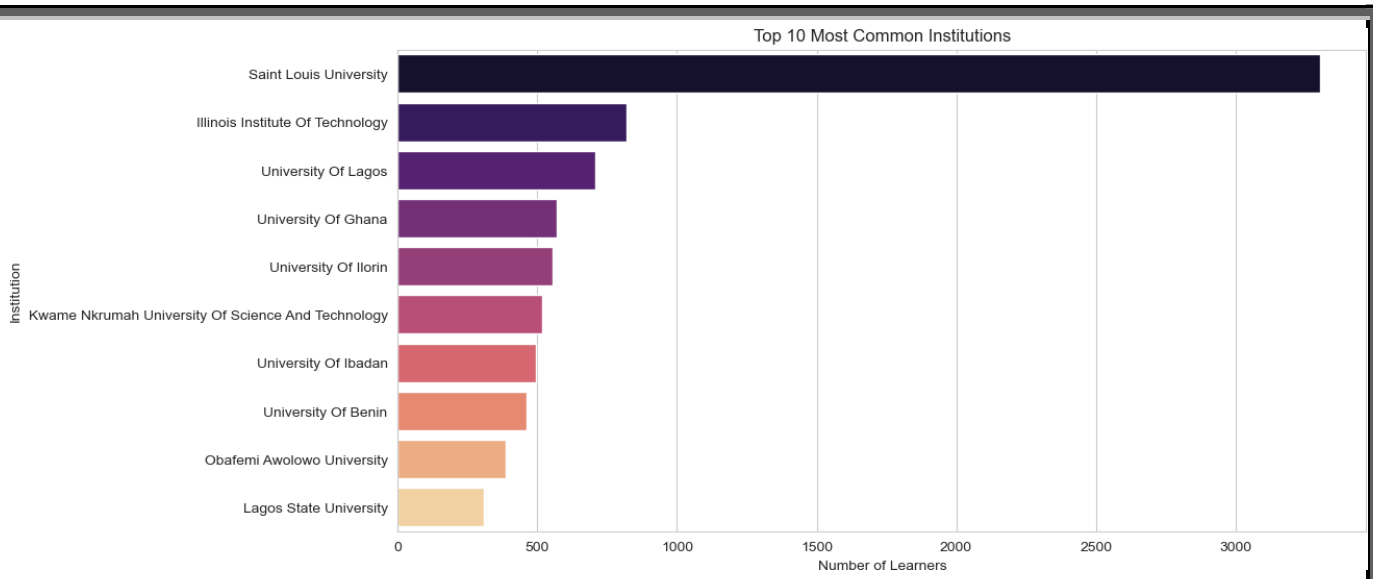
(B) Degree Distribution

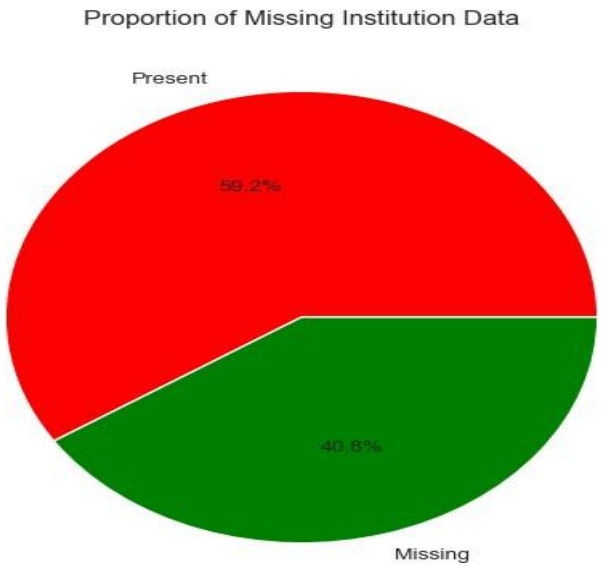
- Degrees have only **8 unique values**, with "Null" being the most common.
- A countplot could visualize the distribution.



(C) Institution Variability

- Over **27,000 unique institutions**.
- The presence of "Null" suggests inconsistencies in data entry.





5) Key findings and next steps for data cleaning and transformation:

Key Findings:

- High missing/null values in key fields.
- Country data mismatch between datasets.
- Large variety of institutions, requiring potential standardization.

Next Steps for Data Cleaning & Transformation:

- Replace "Null" with proper NULL values.
- Investigate and fill missing `Country` values where possible.

- Standardize column names (`Major` vs. `Degree.1`).
- Normalize institution names to avoid duplicates (e.g., "MIT" vs. "Massachusetts Institute of Technology").
- Perform deeper analysis post-cleaning

6. Learner Opportunity Data (learner_opportunity_raw.csv)

By **Areeba Fatima (areebafatima721@gmail.com)**

1) Overview of the dataset structure, sources, and key attributes:

- Data Name: Learner Opportunity Data

This dataset tracks learners' participation in different opportunities, linking user enrollments to programs, including:

- enrollment_id
- learner_id
- assigned_cohort
- apply_date
- status

- Key Attributes:

This dataset contains 113602 rows with 05 columns. Understanding this dataset allows us to uncover trends and outcomes.

ATTRIBUTES	DATATYPE
enrollment_id	VARCHAR
learner_id	VARCHAR
assigned_cohort	VARCHAR
apply_date	TIMESTAM
status	VARCHAR

- Data Source
- Origin: Internal data source
- Frequency of Updates: Historical data

2) **Summary statistics of key variables:**

Below are key summary statistics of the dataset:

ATTRIBUTE	DATATYPE	Start date	End date
Apply_date	TIMESTAMP	2022-06-09 16:28:33.977	2025-02-25 05:15:42.257

3) **Identification of missing values, duplicates, and inconsistencies:**

- Missing Values
  - **Missing values** were detected in the assigned_cohort and status columns.
- Duplicates
  - **No Duplicates were found in the dataset**.
- Fixing Inconsistencies

Changing general datatype to timestamp datatype

1. **Select the column** apply_date.
2. Go to the **Home** tab and click on **Number Format** (dropdown in the "Number" section).
3. Click **More Number Formats**.
4. In the **Format Cells** window, select **Custom**.
5. Enter the format: **YYYY-MM-DD HH:MM:SS**
6. Click **OK**.
7. Copy the new column and **Paste as Values** (**Ctrl + Alt + V → Select "Values"**).

This ensures that the format is actually changed when saving as a CSV.

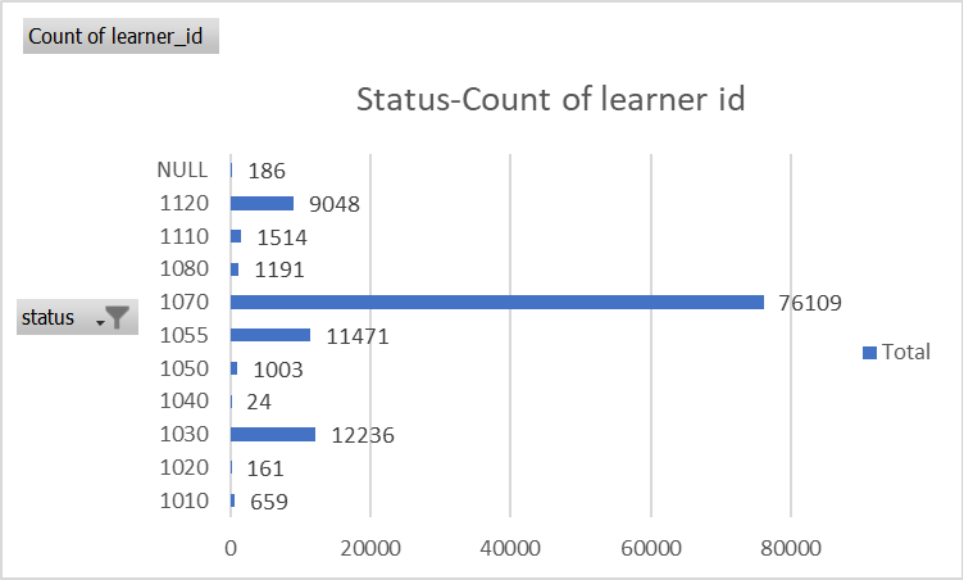
Steps to Clean and Format assigned_cohort in Excel

1. Remove Extra Spaces – Use =TRIM(C2) to remove leading, trailing, and extra spaces between words.
2. Replace Original Data – Copy and Paste Special → Values to overwrite original columns.

4) Data visualizations:

- Visualizations used to understand the dataset:
- Bar chart shows the relationship between the status and learner_id.

Row Labels	Count of learner_id
1010	659
1020	161
1030	12236
1040	24
1050	1003
1055	11471
1070	76109
1080	1191
1110	1514
1120	9048
NULL	186
Grand Total	**113602**



5) Key findings and next steps for data cleaning and transformation:

Key Findings

- **No duplicate** values were detected in the dataset.

- The dataset is ready for further cleaning and transformation before visualization.

Next Steps

- Proceed with **data cleaning** to ensure accurate analysis.
- Prepare the cleaned dataset for transformation and visualization.

7. Cognito Data (cognito_raw.csv)

By **Areeba Fatima (areebafatima721@gmail.com)**

1) **Overview of the dataset structure, sources, and key attributes:**

- Data Name: Cognito Data

The dataset used in this analysis, **Cognito Data (cognito_raw.csv)** – This dataset contains authentication and profile metadata, including:

- Email
- Gender
- Location details

- Key Attributes:

This dataset contains 129178 rows with 9 columns. Understanding this dataset allows us to uncover trends and outcomes.

ATTRIBUTES	DATATYPE
User_id	VARCHAR
Email	VARCHAR
Gender	VARCHAR
UserCreateDate	TIMESTAM
UserLastModifiedDa	TIMESTAM
Birthdate	DATE
City	VARCHAR
Zip	VARCHAR
State	VARCHAR

- *Data Source*
  - Origin: Internal data source
  - Frequency of Updates: Historical data

2) **Summary statistics of key variables:**

Below are key summary statistics of the dataset:

ATTRIBUTE	DATATYPE	Start date	End date
UserCreateDate	TIMESTAMP	2023-01-05 16:32:30.99	2025-02-25 00:34:25.207
UserLastModifiedDate	TIMESTAMP	2023-01-05 19:03:50.665	2025-02-25 00:34:25.207

3) **Identification of missing values, duplicates, and inconsistencies:**

- Missing Values
  - **Missing values** were detected in the Gender, city, birthdate, state & zip columns.
- Duplicates
  - **09 Duplicates were found in email column**.

- Fixing Inconsistencies
  - Changing general datatype to timestamp datatype

1. **Select the column** UserCreateDate.
2. Go to the **Home** tab and click on **Number Format** (dropdown in the "Number" section).
3. Click **More Number Formats**.
4. In the **Format Cells** window, select **Custom**.
5. Enter the format: **YYYY-MM-DD HH:MM:SS**
6. Click **OK**.
7. Copy the new column and **Paste as Values** (**Ctrl + Alt + V → Select "Values"**).

This ensures that the format is actually changed when saving as a CSV.

Same goes for column Userlastmodifieddate.

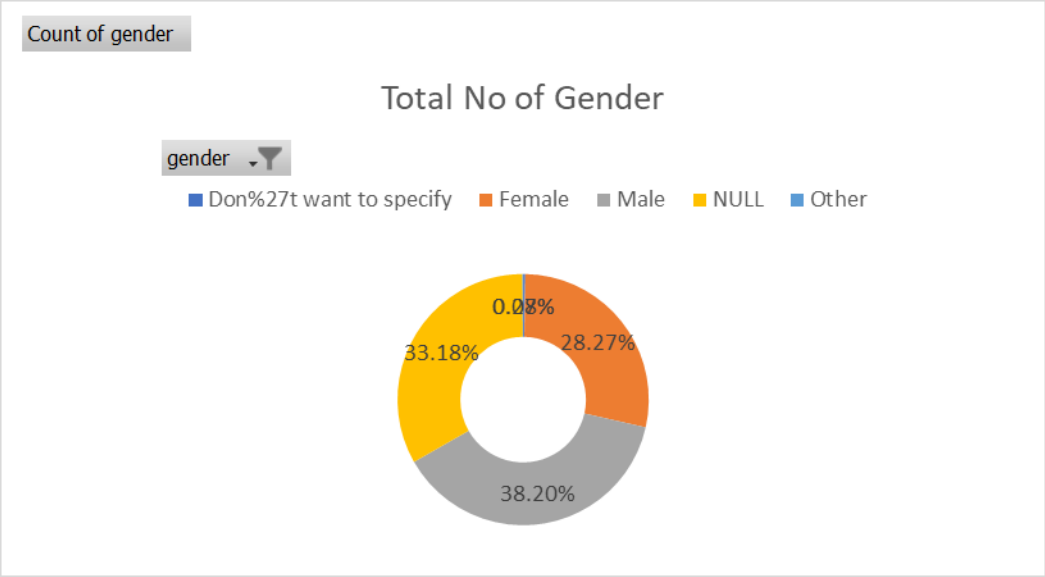
Steps to Clean and Format City & State Columns in Excel

- Remove Extra Spaces – Use =TRIM(G2) to remove leading, trailing, and extra spaces between words.
- Convert to Uppercase – Use =UPPER(TRIM(G2)) to capitalize text.
- Replace Original Data – Copy and Paste Special → Values to overwrite original columns.

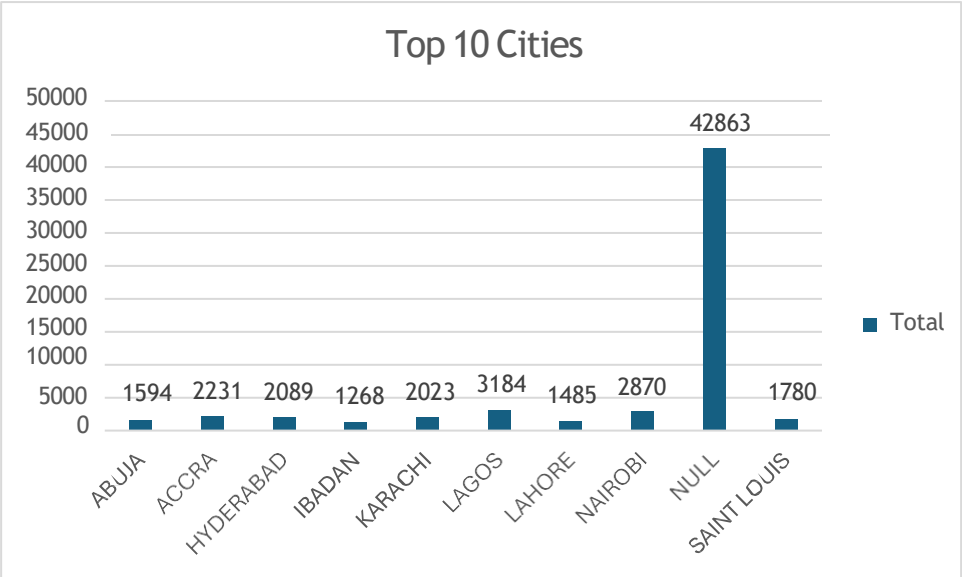
4) **Data visualizations:**

- Visualizations used to understand the dataset:
- Pie chart shows the total percentage of gender in Cognito dataset

Row Labels	Count of gender
Don%27t want to specify	0.28%
Female	28.27%
Male	38.20%
NULL	33.18%
Other	0.07%
Grand Total	**100.00%**



- Bar chart shows the top 10 cities as per Cognito dataset.



5) **Key findings and next steps for data cleaning and transformation:**

- Key Findings
  - **09 duplicate** values were detected in email column.
  - The dataset is ready for further cleaning and transformation before visualization.
- Next Steps
  - Proceed with **data cleaning** to ensure accurate analysis.
  - Prepare the cleaned dataset for transformation and visualization.

**********Thank you**********