Internship Report

Data Visualization and Analyzation
Week - 1

Date : 09-06-2025

Team Number : 26

Team Members:

1. Khushi Khati
(khushikhati11@gmail.com)

2. Precious Okonkwo
(Preciousokonkwo300@gmail.com)

3. Rohit Emmanuel
(kingrohit2439@gmail.com)

4. Shomitra Dey
(soumitradev532@gmail.com}

Abstract

This EDA and visualization report presents a comprehensive exploratory data analysis of six datasets—User Data, Opportunity Data, Cohort Data, Marketing Data, Learner Opportunity Data, and Cognito Data—pertaining to a learning platform's operations. The report provides an overview of the dataset structure and key attributes, including column details and data types, alongside summary statistics of key variables. It identifies the presence of missing values and duplicates within each dataset, offering insights into data quality without any manipulation or processing of the raw data. The analysis is enriched with a variety of visualizations, such as histograms, boxplots, pie charts, bar charts, and tables, to illustrate distributions, trends, and anomalies. Serving as a foundational document, this report establishes a baseline for future work, including data cleaning, preprocessing, and advanced analytical endeavors, thereby facilitating informed decision-making and further exploration of the platform's dynamics.

Contents

Chapter 1

Introduction

This EDA and visualization report is designed to delve deeply into the analysis and visualization of six distinct datasets, aiming to uncover valuable insights into user behavior, participation in learning programs, and the effectiveness of marketing strategies. The datasets under examination include User Data, Opportunity Data, Cohort Data, Marketing Data, Learner Opportunity Data, and Cognito Data, each serving a critical role in understanding the multifaceted operations of the learning platform. Each dataset has been subjected to a comprehensive exploratory data analysis (EDA) process, enriched with a variety of visualizations—such as histograms, boxplots, and pie charts—to consolidate findings and provide a thorough understanding of the platform's performance metrics and user interactions.

This report meticulously documents the completed analyses for all six datasets, offering detailed insights into user demographics, program engagement, cohort trends, marketing performance, learner participation, and the quality of authentication metadata. The User Data dataset, for instance, reveals enrollment patterns and demographic distributions, while Opportunity Data highlights program popularity and engagement metrics. Cohort Data provides a window into participation trends, Marketing Data assesses campaign efficacy, Learner Opportunity Data tracks user engagement across programs, and Cognito Data exposes data quality issues like missing values. These analyses are conducted on raw, unprocessed data, ensuring an authentic baseline for further investigation.

The significance of this work lies in its establishment of a robust foundation for advanced analysis and transformation of the cleaned datasets in the weeks to follow. By identifying key variables, detecting anomalies, and visualizing trends, this report enables informed decision-making and sets the stage for data cleaning, preprocessing, and the development of predictive models. The insights derived will support strategic enhancements to the learning platform, addressing user needs, optimizing marketing efforts, and improving data integrity. As such, this report not only encapsulates the initial phase of the project but also paves the way for a deeper, more refined exploration of the platform's dynamics, fostering a data-driven approach to future enhancements and operational improvements.

Chapter 2

Dataset Overview

The project involves six datasets, each offering unique insights into the learning platform's operations. Below is a summary of each dataset, based on the exploratory data analysis (EDA) conducted in Chapter 3.

2.1 User Data (user_data.csv)

The User Data dataset, stored in a PostgreSQL database and sourced from 'user_data.csv', contains 129,259 records with five columns: 'learner_id', 'country', 'degree', 'institution', and 'major', all of type 'character varying'. It captures learner demographics and educational background, enabling analysis of enrollment trends and user characteristics through demographic distributions and temporal patterns.

2.2 Opportunity Data (opp_data.csv)

The Opportunity Data dataset, sourced from 'opp_data.csv', provides details about learning opportunities, including program descriptions, cohort associations, sponsorships, and participation metrics. It supports analysis of program popularity and engagement, with visualizations highlighting the distribution of opportunities across categories and their types.

2.3 Cohort Data (cohort_data.csv)

The Cohort Data dataset, sourced from 'cohort_data.csv', contains 639 rows and 5 columns, tracking cohort-based learning programs with attributes like 'cohort_id', 'cohort_code', 'start_date', 'end_date', and 'size'. It allows analysis of participation and completion trends, with a right-skewed distribution of cohort sizes and outliers identified via boxplots.

2.4 Marketing Data (marketing_data.csv)

The Marketing Data dataset, sourced from 'marketing_data.csv', captures advertising performance metrics, including campaign reach, engagement, and costs. It enables evaluation of marketing effectiveness in driving enrollments, with analyses revealing the re-

lationship between ‘cost_per_result‘ and ‘results‘, and distributions of ‘ad_account_name‘ and ‘result_type‘.

2.5 Learner Opportunity Data (learner_opportunity_raw.csv)

The Learner Opportunity Data dataset, sourced from ‘learner_opportunity_raw.csv‘, comprises 113,602 rows and 5 columns, tracking learners’ participation with attributes like ‘enrollment_id‘, ‘learner_id‘, ‘assigned_cohort‘, ‘apply_date‘, and ‘status‘. It supports analysis of user engagement, with insights into application trends and status distributions, including outlier detection in ‘status‘.

2.6 Cognito Data (cognito_raw.csv)

The Cognito Data dataset, sourced from ‘cognito_raw.csv‘, contains 129,178 records with authentication and profile metadata, including ‘user_id‘, ‘email‘, ‘gender‘, ‘birthdate‘, ‘city‘, ‘zip‘, ‘state‘, ‘user_create_date‘, and ‘user_last_modified_date‘. EDA highlighted high missing values in ‘gender‘, ‘city‘, ‘zip‘, and ‘state‘, with visualizations of gender distribution and city demographics.

Chapter 3

Exploratory Data Analysis and Visualizations

3.1 User Data

The User Data dataset, stored in a PostgreSQL database, contains 129,259 records with five columns: `learner_id`, `country`, `degree`, `institution`, and `major`, all of type `character varying`. The dataset, sourced from `user_data.csv`, captures learner demographics and educational background.

3.1.1 Dataset Structure and Summary Statistics

The dataset structure was examined to understand its composition. A preview of the first 10 rows (Figure 3.1) illustrates the diversity of learner profiles, including countries like Nigeria, Kenya, Bangladesh, and Pakistan, and degrees ranging from Undergraduate to Graduate Students.

	learner_id character varying (200)	country character varying (200)	degree character varying (200)	institution character varying (200)	major character varying (200)
1	Learner#00004f18-8b86-4fe4-ad7e-6c8d988f5335	Nigeria	Undergraduate Student	Federal University of Technology Owerru	Civil Engineering
2	Learner#00006478-745f-49bf-b126-02584e830720	Nigeria	NULL	NULL	NULL
3	Learner#00010567-1336-433c-a941-a612b3d2fbb8	Kenya	Graduate Student	UNICAF UNIVERSITY	Environmental Sustainabi
4	Learner#00011c80-0c5c-4601-9696-b3ca787e264f	Bangladesh	NULL	NULL	NULL
5	Learner#000141a7-4c82-401f-a2e6-dd12b4b26089	Nigeria	NULL	NULL	NULL
6	Learner#0acf3501-57a8-4585-91e3-6bbb89d3c801	Ghana	NULL	NULL	NULL
7	Learner#0001ca2c-7bec-4a33-833c-b844a29f4dea	Nigeria	Graduate Student	Nasarawa State University, Keffi	Accounting
8	Learner#0003bed9-d9d9-49a7-a755-a9562aaa0dfb	Pakistan	Graduate Student	CTTI College KP Campus	Part Time
9	Learner#0004295c-717e-4953-b3bf-fff2daf0e903	Nigeria	Undergraduate Student	Caleb University	Computer Science
10	Learner#00049a81-94a9-4b25-92ed-62d017f3b67a	Philippines	Undergraduate Student	Our lady of fatima university	Nursing

Figure 3.1: Preview of the first 10 rows of the User Data dataset.

The column names and data types (Figure 3.2) confirm that all columns are of type `character varying`, suitable for categorical and textual data.

Summary statistics revealed 191 unique countries, with India and Nigeria being the most frequent. The `degree` column showed 52,693 (40.77%) missing values, with Graduate Students (31,806) and Undergraduate Students (30,709) as the most common categories among valid entries. The `major` column listed varied fields like Computer Science, Civil Engineering, and Nursing, with 52,697 (40.77%) missing values. The top 10 majors in-

	column_name name	data_type character varying
1	learner_id	character varying
2	country	character varying
3	degree	character varying
4	institution	character varying
5	major	character varying

Figure 3.2: Column names and data types of the User Data dataset.

cluded Computer Science, Engineering, and Business Administration, indicating a tech and business focus.

3.1.2 Data Visualizations

- **Pie Chart: User Distribution by Country** (Figure 3.3): Shows India and Nigeria as dominant, with smaller contributions from Pakistan, Kenya, the United States, and Ghana, highlighting South Asian and West African representation.

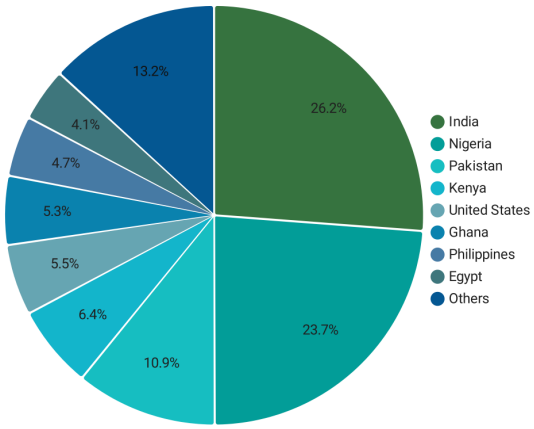

Figure 3.3: Pie chart of user distribution by country in User Data.

- **Bar Chart: User Distribution by Institution** (Figure 3.4): Displays top institutions like Saint Louis University and University of Lagos, with 40.9% missing data.

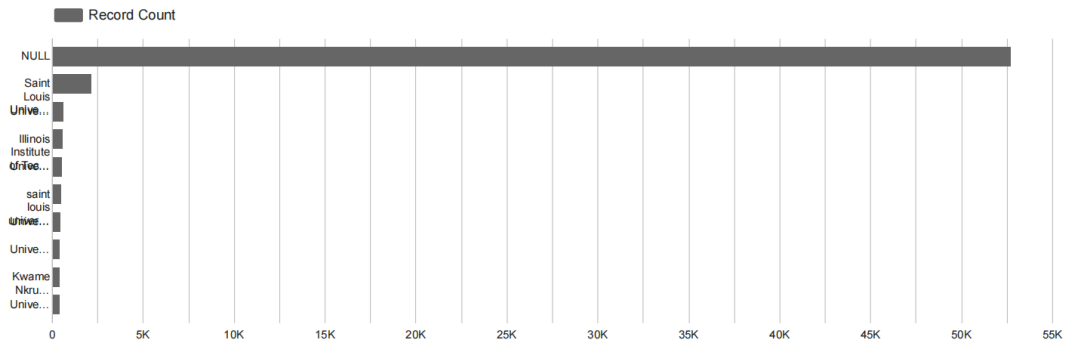

Figure 3.4: Bar chart of user distribution by institution in User Data.

- **Bar Chart: User Distribution by Degree** (Figure 3.5): Indicates Graduate Students slightly outnumber Undergraduates, with 40.77% missing data.

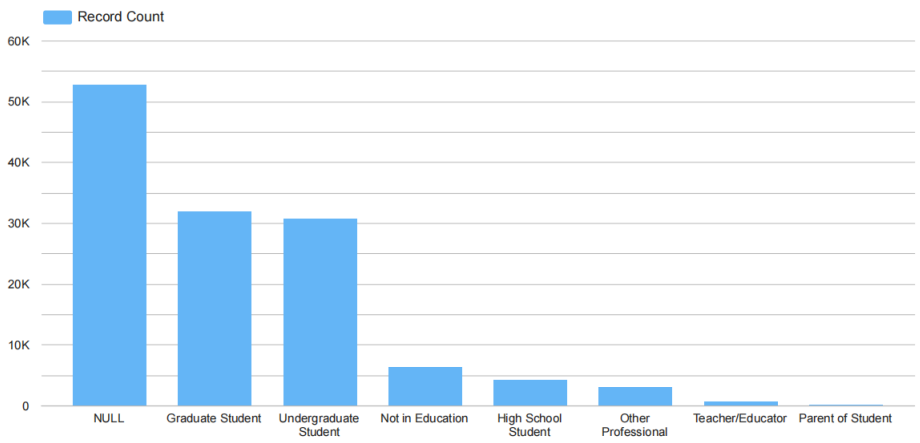

Figure 3.5: Bar chart of user distribution by degree in User Data.

- **Horizontal Bar Chart: User Distribution by Major** (Figure 3.6): Highlights Computer Science, Engineering, and Business Administration as top majors, with 40.77% missing data.

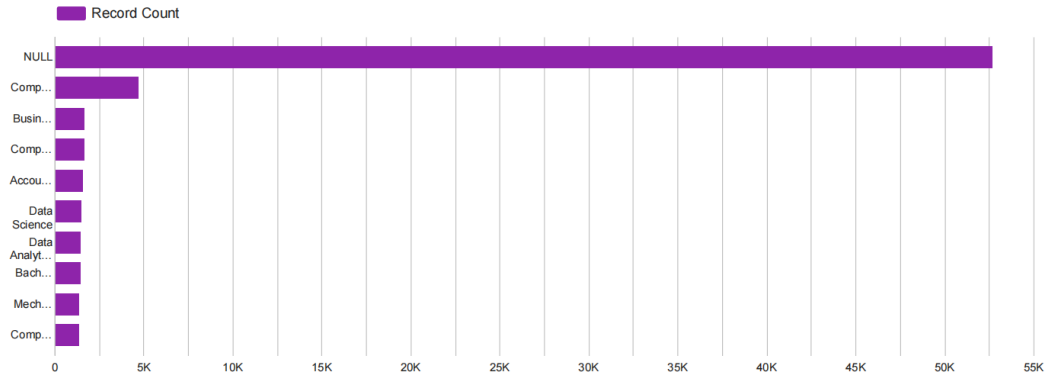

Figure 3.6: Horizontal bar chart of top 10 majors in User Data.

3.1.3 Missing Values, Duplicates, and Inconsistencies

Missing values were significant: `country` (2,275, 1.76%), `degree` (52,693, 40.77%), `institution` (52,901, 40.9%), and `major` (52,697, 40.77%), with `learner_id` having none. No duplicate records were found, ensuring data uniqueness. No numerical columns were present for outlier detection.

3.2 Opportunity Data

The Opportunity Data dataset, sourced from ‘opp_data.csv‘, provides details about learning opportunities, including program descriptions, cohort associations, sponsorships, and participation metrics. It supports analysis of program popularity and engagement.

3.2.1 Dataset Structure and Summary Statistics

The dataset structure was explored by querying the data types of columns using the ‘information_schema.columns‘ table, revealing the schema of the ‘opportunity_data‘ table. A preview of the first few rows was obtained to illustrate the dataset’s composition (see Figure 3.7). The total number of rows was determined to be [value pending], providing an overview of the dataset’s size.

	opportunity_id text	opportunity_name text	category text	opportunity_code text	tracking_questions text
1	Opportunity#000000000G127E8VYE08TXBT6X	Choosing and Planning for Your Major	Event	E501873	[null]
2	Opportunity#000000000G2PB6VB4ANR28CV2P	The Financial: Article Writing Competition Test	Competiti...	M523594	[null]
3	Opportunity#000000000G44AM4J9NBMPK8TJ...	Entrepreneurship and Innovation	Internship	I289641	[null]
4	Opportunity#000000000G4F19XBEXPWKS8F3N	Statement of Purpose (SOP) Writing Workshop	Event	E258709	[null]
5	Opportunity#000000000G4KVEP36NNJR5YWTJ	Project Management	Internship	I584159	[null]
6	Opportunity#000000000G8BW90E86ARRKM3...	Cybersecurity: Defensive Hacking	Internship	I155449	[null]
7	Opportunity#000000000G8JG2FEA12SVNXXEN	Esports and Game Design	Internship	I860340	[null]
8	Opportunity#000000000G95BD07NB0181K0XD	Data Visualization	Internship	I660879	[null]
9	Opportunity#000000000GBZ5VRTC3YS9T716N	Data Visualization	Internship	I755008	[null]
10	Opportunity#000000000GCKFV5K6Q8FWGFH...	Choosing and Planning for Your Major + Career Exploration Workshops ? In-person	Event	E189919	[null]
11	Opportunity#000000000GCTJ4F7QXJWWM8D...	Major and Career Exploration Workshop	Event	E352968	[null]
12	Opportunity#000000000GDD59YDSJCXA2H46X	Entrepreneurship and Innovation	Internship	I252028	[null]
13	Opportunity#000000000GEHAHYHGRSY59TR...	Building a Strong Application in a Test-Optional World	Event	E322161	[null]
14	Opportunity#000000000GGG8B9VDBKAQM1D9...	Teaching During The Time Of Disruption Test	Event	E289640	[null]
15	Opportunity#000000000GGJJG260YZ8XEWZV...	Digital Marketing	Internship	I127099	[null]
16	Opportunity#000000000GH4BAHFS8NTZC0489	Million Dollar Idea	Competiti...	M584683	[null]
17	Opportunity#000000000GHGN3NGS84M44MK...	Slide Geeks: A Presentation Design Competition	Competiti...	M429108	[null]
18	Opportunity#000000000GHXFZ9DEH0TDCZPF5	Junior Pumpkin Launch ? In-person	Competiti...	M103449	[null]
19	Opportunity#000000000GKZ7RSWWZA14JDC...	Major and Career Exploration Workshop ? Virtual	Event	E296277	[null]
20	Opportunity#000000000GM6HGM7M067WVT...	Virtual Internship Facilitator Test	Career	A933703	[null]
21	Opportunity#000000000GN2A0AY7XK8CSFZPP	Career Essentials: Getting Started with Your Professional Journey	Course	U533812	{is_required_for_badge_award:true,code:QBWTGGT,question:All deliverables submitted,is_frozen:false,ans_type:boolean}

Figure 3.7: Preview of the first few rows of the Opportunity Data dataset.

	column_name name	data_type character varying
1	opportunity_id	text
2	opportunity_name	text
3	category	text
4	opportunity_code	text
5	tracking_questio...	text

Figure 3.8: Column names and data types of the Opportunity Data dataset.

3.2.2 Visualizations

A bar chart visualizes the distribution of opportunities across different categories (see Figure 3.9). Additionally, a pie chart highlights the proportion of Internship, Event, Competition, and Career opportunities relative to all others (see Figure 3.10).

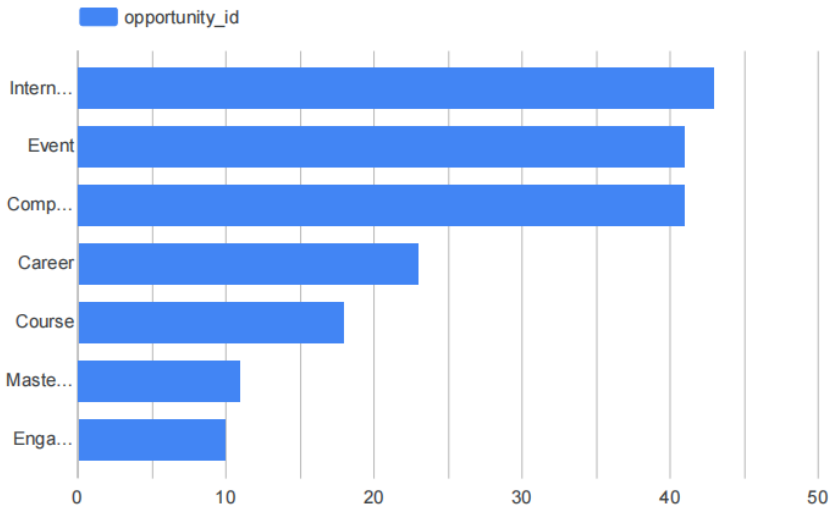

Figure 3.9: Bar chart of opportunity distribution across categories in Opportunity Data.

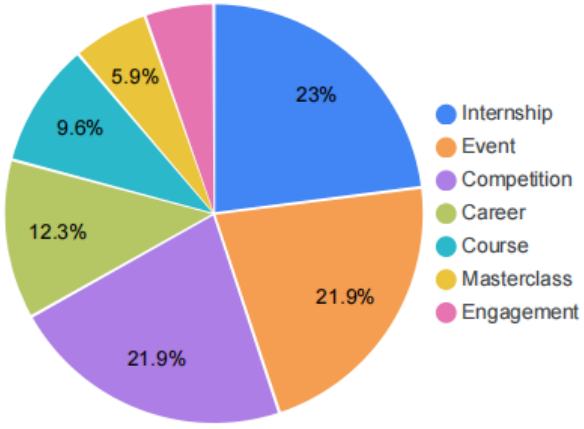

Figure 3.10: Pie chart of opportunity type proportions (Internship, Event, Competition, Career) in Opportunity Data.

3.2.3 Missing Values, Duplicates, and Inconsistencies

Missing values were assessed across the columns of the Opportunity Data dataset: ‘opportunity_id‘ (0), ‘opportunity_name‘ (0), ‘category‘ (0), ‘opportunity_code‘ (0), and ‘tracking_questions‘ (69). The exact percentage for ‘tracking_questions‘ is pending due to the unknown total number of rows, but it indicates a small proportion of missing data in this column. No duplicate records were identified, as the duplicate check query grouping all columns (‘opportunity_id‘, ‘opportunity_name‘, ‘category‘, ‘opportunity_code‘, ‘tracking_questions‘) with a ‘HAVING COUNT(*) > 1‘ condition yielded no results. No numerical columns were present for outlier detection.

3.3 Cohort Data

The Cohort Data dataset, sourced from ‘cohort_data.csv‘, contains 639 rows and 5 columns, tracking cohort-based learning programs. The key attributes are:

- `cohort_id`: Identifier for the cohort (1 unique value).
- `cohort_code`: Code representing the cohort (639 unique values).
- `start_date`: The start date of the cohort (datetime format).
- `end_date`: The end date of the cohort (datetime format).
- `size`: The size of the cohort (numerical).

3.3.1 Dataset Structure and Summary Statistics

Summary statistics for the numeric column `size` are as follows:

Statistic	Size
Count	639
Unique	53
Top	800
Frequency	171
Mean	NaN

There are no missing values in any of the columns, and no duplicate entries were found in the dataset.

3.3.2 Missing Values and Duplicates

No missing values were found in any of the columns. Duplicates were confirmed to be absent based on the relevant keys.

3.3.3 Outliers and Inconsistencies

A boxplot was created to visualize the distribution of cohort sizes, aiding in the identification of any outliers (see Figure 3.11).

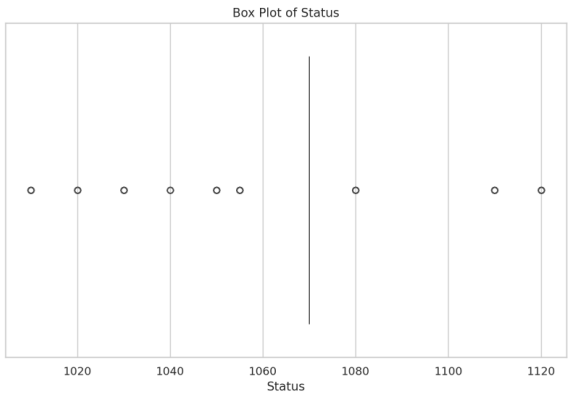

Figure 3.11: Boxplot of cohort sizes in Cohort Data.

3.3.4 Visualizations

The distribution of cohort sizes was visualized using a histogram (see Figure 3.12). Key points include:

- **Bars**: Each bar represents the count of cohorts within a specific size range, with the height indicating the number of cohorts.
- **Distribution Shape**: The histogram is right-skewed, indicating most cohorts have smaller sizes, with a few significantly larger ones suggesting outliers.
- **KDE (Kernel Density Estimate)**: The smooth curve overlaid on the histogram represents the estimated probability density function, enhancing the visualization of the distribution.

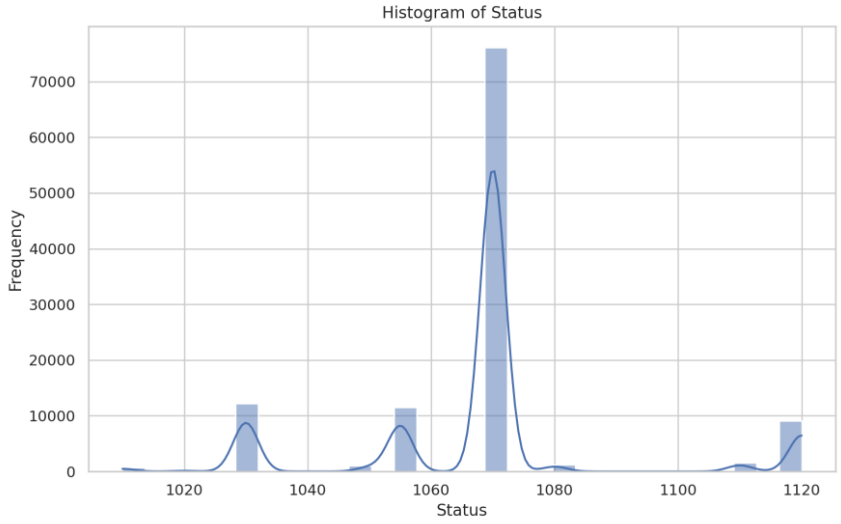

Figure 3.12: Histogram of cohort sizes in Cohort Data with KDE overlay.

3.3.5 Missing Values, Duplicates, and Inconsistencies

Missing values were assessed across the columns of the Cohort Data dataset: ‘cohort_id‘, ‘cohort_code‘, ‘start_date‘, ‘end_date‘, and ‘size‘, with a total of 639 rows. No missing values were found in any of the columns, indicating complete data coverage. No duplicate records were identified, as confirmed by the analysis of relevant keys across all columns. Outlier detection was performed on the numerical column ‘size‘, with a boxplot revealing a right-skewed distribution and potential outliers among larger cohort sizes.

3.4 Marketing Data

The Marketing Data dataset, sourced from ‘marketing_data.csv‘, captures advertising performance metrics, such as campaign reach, engagement, and costs. It enables evaluation of marketing effectiveness in driving enrollments.

3.4.1 Dataset Structure and Summary Statistics

The dataset structure was explored by querying the data types of columns using the ‘information_schema.columns‘ table, revealing the schema of the ‘marketing_campaign‘ table. A preview of the first few rows was obtained to illustrate the dataset’s composition (see Figure 3.13). The total number of rows was determined to be [value pending], providing an overview of the dataset’s size.

	ad_account_name text	campaign_name text	delivery_status text	delivery_level text	reach integer	outbound_clicks integer	landing_page_views integer	result_type text	results integer	cost_per_result double precision	amount_spent_aed double precision	cpc_cost_per_link_click double precision	reporting_starts date
1	SLU	###B2: Digital Marketing Intern - May Ads: Website Leads Prospecting | 18 to 35 years - Copy 4	completed	campaign	102962	1815	1310	Website applications submitt...	386	1.910959	737.63	0.406184	2023-01-01
2	SLU	##B2: Digital Marketing Intern - May Ads: Website Leads Prospecting | 18 to 35 years - Copy 3	completed	campaign	180175	3378	2152	Website applications submitt...	598	1.233495	737.63	0.217976	2023-01-01
3	SLU	#B2: Digital Marketing Intern - May Ads: Website Leads Prospecting | 18 to 35 years - Copy 3	inactive	campaign	173118	3591	2767	Website applications submitt...	514	1.745914	897.4	0.249486	2023-01-01
4	SLU	#Brand Awareness: UGC Video - March - Copy	inactive	campaign	18355415	5431	1019	Reach	18355415	0.092997	1707	0.309744	2023-01-01
5	SLU	#Data Analyst Associate Internship	inactive	campaign	2448	111	62	Website leads	1	5.43	5.43	0.048919	2023-01-01
6	SLU	A1: Outreach Consultant Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	1136229	12798	8196	Website leads	2933	1.175482	3447.69	0.269288	2023-01-01
7	SLU	A2: Digital Strategy Associate Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	682351	9510	6912	Website leads	2945	1.162503	3423.57	0.359392	2023-01-01
8	SLU	A3: Data Analyst Associate Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	1077778	21464	16914	Website leads	8025	0.436908	3506.19	0.162988	2023-01-01
9	SLU	A4: Project Management Associate Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy...	inactive	campaign	934611	13060	8935	Website leads	4540	0.749714	3403.7	0.260281	2023-01-01
10	SLU	A5: Business Strategy Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	999159	13337	9112	Website leads	4113	0.906715	3729.32	0.279057	2023-01-01
11	SLU	AeroFlow Challenge	completed	campaign	142743	1922	1149	Website applications submitt...	324	4.555556	1476	0.765957	2023-01-01
12	SLU	AI Forensic Challenge	completed	campaign	114896	1122	818	Website applications submitt...	249	2.920683	727.25	0.646444	2023-01-01
13	SLU	AUG: AeroFlow Challenge	completed	campaign	64594	804	468	Website applications submitt...	200	5.5095	1101.9	1.370522	2023-01-01
14	SLU	AUG: Power Course: Digital Wellness - Copy	inactive	campaign	169228	660	271	Website applications submitt...	326	3.379202	1101.62	1.669121	2023-01-01
15	SLU	AUG_Course: CPR course - Copy	completed	campaign	148357	2037	1297	Website applications submitt...	396	3.403788	1347.9	0.661384	2023-01-01
16	SLU	Awareness: Do you want to stand out? | Reel and story | Video Views	completed	campaign	1261904	3644	808	ThruPlay	216248	0.00851	1840.17	0.504433	2023-01-01
17	Brand Awareness	AWARENESS: Excelerate_Video_Ad (with Exclusion)	completed	campaign	1560112	1850	15	ThruPlay	178701	0.010277	1836.42	0.988918	2023-01-01
18	SLU	B1: Outreach Consultant Intern - May Ads: Website Leads Prospecting | 18 to 35 years	completed	campaign	275349	5331	3820	Website applications submitt...	815	2.707718	2206.79	0.398769	2023-01-01
19	SLU	B2: Digital Marketing Intern - May Ads: Website Leads Prospecting | 18 to 35 years - Copy 3	inactive	campaign	1315	14	12	Website applications submitt...	3	1.293333	3.88	0.277143	2023-01-01
20	SLU	B3: Data Analyst Associate Intern - May Ads: Website Leads Prospecting | 18 to 35 years - Copy 3	completed	campaign	417944	10165	7708	Website applications submitt...	2169	1.020235	2212.89	0.217355	2023-01-01

Figure 3.13: Preview of the first few rows of the Marketing Data dataset.

	column_name name	data_type character varying
1	ad_account_name	text
2	campaign_name	text
3	delivery_status	text
4	delivery_level	text
5	reach	integer
6	outbound_clicks	integer
7	landing_page_views	integer
8	result_type	text
9	results	integer
10	cost_per_result	double precision
11	amount_spent_aed	double precision
12	cpc_cost_per_link_cli...	double precision
13	reporting_starts	date

Figure 3.14: Column names and data types of the Marketing Data dataset.

3.4.2 Visualizations

A table visualizes the relationship between ‘cost_per_result‘ and ‘results‘, showing that achieving more results typically costs less per result, with some campaigns being highly efficient and others less so (see Figure 3.15). A pie chart displays the proportion of campaigns per ‘ad_account_name‘ based on record count (see Figure 3.16). A bar chart illustrates the distribution of ‘result_type‘ versus record count, highlighting the primary goals of campaigns (see Figure 3.17). A donut chart shows the proportion of campaigns for each ‘result_type‘, with "Website applications submitted" as the largest segment (see Figure 3.18).

	cost_per_result ▾	results
1.	47.08949	78
2.	14.93333	3
3.	13.33032	124
4.	13.20011	1,004
5.	12.74326	144
6.	11.15944	18
7.	11.08536	166
8.	10.9001	202
9.	10.77667	33
10.	10.68105	19

1 - 100 / 141

Figure 3.15: Table of cost_per_result vs. results in Marketing Data.

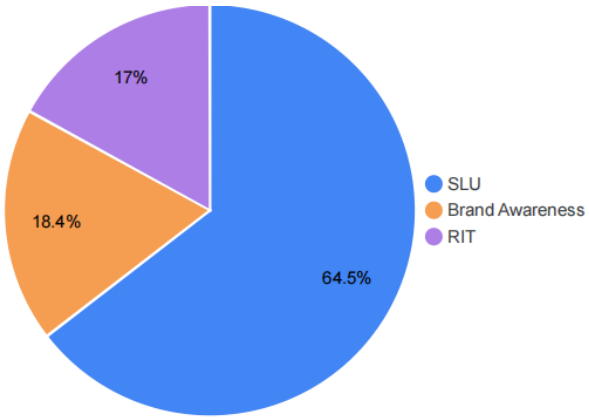

Figure 3.16: Pie chart of campaign proportions per ad account name in Marketing Data.

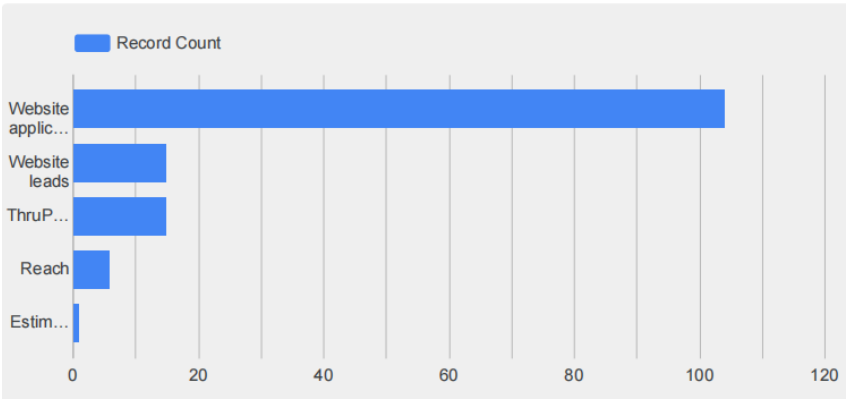

Figure 3.17: Bar chart of result type distribution in Marketing Data.

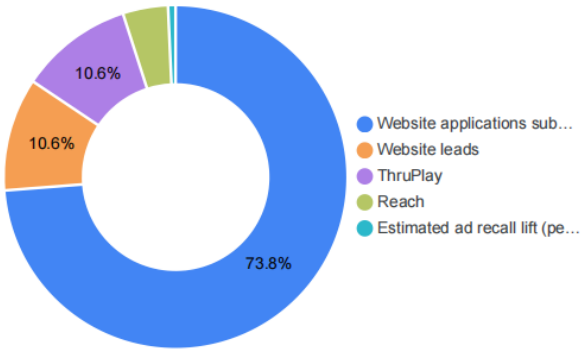

Figure 3.18: Donut chart of result type proportions in Marketing Data.

3.4.3 Identification of Missing Values, Duplicates, and Inconsistencies

Missing value counts were calculated for columns including ‘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_cos_per_link_click‘, and ‘reporting_starts‘, identifying the extent of missing data. No duplicate entries were found, as confirmed by grouping all columns and checking for records with a count greater than 1.

3.5 Learner Opportunity Data

The Learner Opportunity Data dataset, sourced from ‘learner_opportunity_raw.csv‘, comprises 113,602 rows and 5 columns, tracking learners’ participation in opportunities. The key attributes are:

- `enrollment_id`: Unique identifier for each enrollment (object type).
- `learner_id`: Unique identifier for each learner (object type).
- `assigned_cohort`: Identifier for the cohort assigned to the learner (object type).
- `apply_date`: Date when the application was submitted (converted to datetime type for better analysis).
- `status`: Current status of the enrollment (float64 type).

3.5.1 Dataset Structure and Summary Statistics

Summary statistics for the key variables are as follows:

- **Numeric Summary**: The `status` column has a mean of approximately 1068.19, with a standard deviation of about 21.03. The minimum value is 1010, and the maximum is 1120.
- **Categorical Summary**: The `learner_id` has 187 unique values, indicating multiple enrollments per learner. The `assigned_cohort` has 575 unique values, with the most frequent cohort being ’Opportunity000000000GWAQAXC5X45C2MHJ28’, appearing 10,772 times.

- **Date Range**: The `apply_date` ranges from June 9, 2022 to February 25, 2025.

3.5.2 Outlier Detection

A boxplot was generated to visualize potential outliers in the `status` column (see Figure 3.19). The boxplot indicates the presence of several outliers, which should be further investigated to assess their validity.

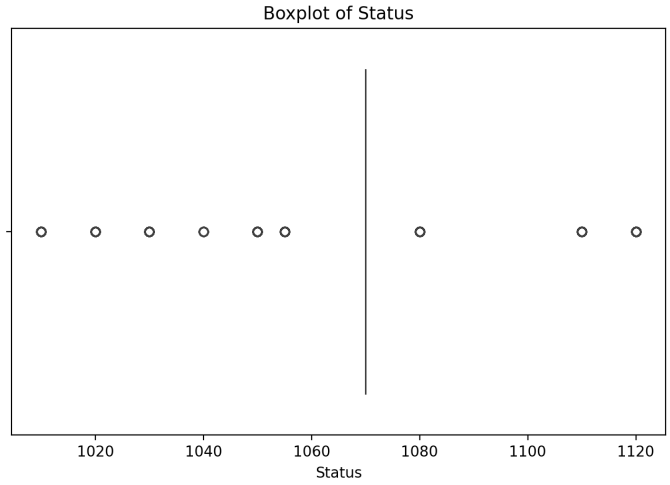

Figure 3.19: Boxplot of status values in Learner Opportunity Data.

3.5.3 Application Trends Over Time

The trend of application submissions over time is visualized below (see Figure 3.20). The chart reveals distinct application spikes in May 2023 and April 2024, suggesting enrollment campaigns or significant events, with a steady rise observed from mid-2024.

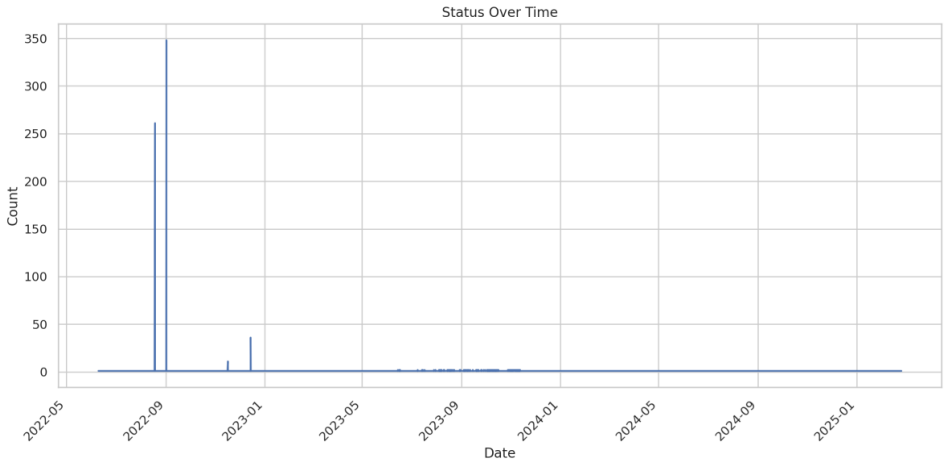

Figure 3.20: Trend of application submissions over time in Learner Opportunity Data.

3.5.4 Distribution of Enrollment Status

A bar plot shows the distribution of values in the `status` column (see Figure 3.21). The majority of records are concentrated around a `status` value of 1070, indicating a dominant enrollment state, with other status values less frequent but consistently present.

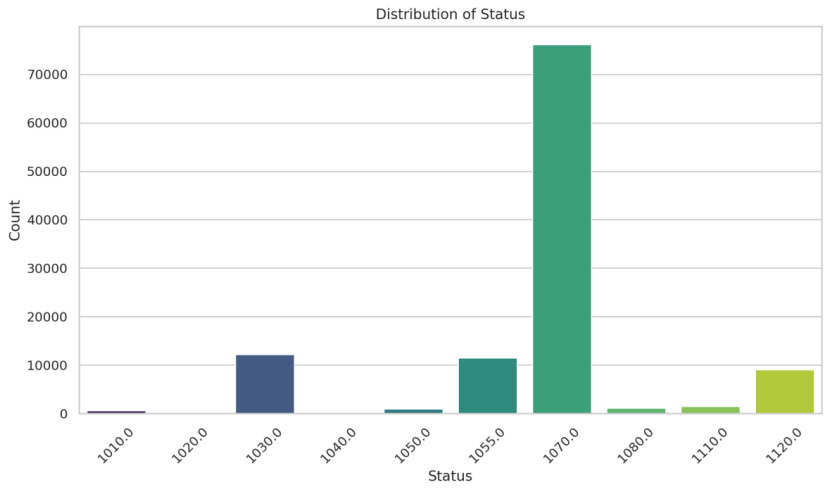

Figure 3.21: Distribution of status values in Learner Opportunity Data.

3.5.5 Key Findings and Next Steps

- **Missing Data Patterns**: The high volume of missing `assigned_cohort` values should be addressed through imputation or by examining contextual patterns.
- **Outliers**: Several outliers were detected in the `status` column, requiring validation or possible exclusion.
- **Temporal Insights**: Sudden spikes in applications may relate to external events or initiatives, which could yield insights for campaign planning.

Next Steps:

- Address missing values with appropriate strategies.
- Analyze and possibly clean outliers.
- Deepen cohort analysis and segment learners by status for further insights.
- Build predictive or diagnostic models based on cleaned data.

3.5.6 Missing Values, Duplicates, and Inconsistencies

Missing values were assessed across the columns of the Learner Opportunity Data dataset, comprising 113,602 rows: ‘enrollment_id‘ (0), ‘learner_id‘ (0), ‘assigned_cohort‘ (13,318, 11.73%), ‘apply_date‘ (188, 0.17%), and ‘status‘ (186, 0.16%). The percentages reflect the proportion of missing values relative to the total rows. No duplicate records were identified, as confirmed by the analysis indicating no duplicate entries in the dataset. Outlier detection was performed on the numerical column ‘status‘, with a boxplot revealing several outliers that require further investigation for validity.

3.6 Cognito Data

The Cognito Data dataset, sourced from ‘cognito_raw.csv‘, contains authentication and profile metadata for users. EDA focused on data quality and key attributes, with analyses performed using SQL queries on the ‘public.users‘ table.

3.6.1 Dataset Structure and Summary Statistics

The dataset structure was explored by retrieving the first few rows using a query to display the initial 100 records ordered by ‘user_id‘. A preview of these rows illustrates the dataset’s composition (see Figure 3.22). Column names and data types were checked to confirm the schema, revealing the structure of the dataset (see Figure 3.23). The total number of rows was determined to be 129,178. Temporal analyses identified the earliest and latest user creation dates (ranging from January 5, 2023, to February 25, 2025), modification dates (ranging from January 6, 2023, to February 25, 2025), and birthdates (ranging from June 19, 1924, to May 27, 2022) for non-null entries.

	user_id [PK] character varying (36)	email character varying (255)	gender character varying (50)	user_create_date timestamp with time zone	user_last_modified_date timestamp with time zone	birthdate character varying (10)	city character varying (100)	zip character varying (20)	state character varying (100)
1	00004f18-8b86-4fe4-ad7e-6c8d988f5335	lilbrianholy@gmail.com	Male	2023-07-23 14:05:58.44+06	2023-07-23 14:09:52.887+06	6/23/2001	Owerri	460103	Imo
2	00006478-745f-49bf-b126-02584e830720	adeyoyintimileyin@gmail.com	[null]	2024-03-30 15:36:30.53+06	2024-03-30 15:37:06.851+06	[null]	[null]	[null]	[null]
3	00010567-1336-433c-a941-a612b3d2fbb8	gikonyosalome19@gmail.com	Female	2024-11-18 03:25:56.381+06	2024-11-18 03:32:50.783+06	5/4/1996	NAIVASHA	20117	NAKURU
4	00011c80-0c5c-4601-9696-b3ca787e264f	shakibsm32@gmail.com	[null]	2024-12-24 14:54:35.686+06	2024-12-24 14:54:45.61+06	[null]	[null]	[null]	[null]
5	000141a7-4c82-401f-a2e6-dd12b4b26089	mandcastle123@gmail.com	[null]	2024-03-31 22:15:13.872+06	2024-03-31 22:21:28.939+06	[null]	[null]	[null]	[null]
6	0001ca2c-7bec-4a33-833c-b844a29f4dea	samemma07@gmail.com	Male	2024-11-13 03:28:48.072+06	2024-11-13 03:36:22.897+06	6/6/1986	Abuja	900211	Abuja FCT
7	0003bed9-d9d9-49a7-a755-a9562aaa0dfb	survival4426@gmail.com	Male	2025-02-12 10:37:48.694+06	2025-02-12 10:44:07.951+06	4/12/1999	Khanur	64100	PUNJAB
8	0003dac8-45a1-40d0-9c9b-9c3cfb7060cf	atiffiroz628@gmail.com	Male	2025-02-21 20:59:07.273+06	2025-02-21 21:03:39.306+06	6/15/2006	Kanpur	208010	Uttarpradesh
9	0004295c-717e-4953-b3bf-fff2daf0e903	ogunleyetunde.r@gmail.com	Male	2024-04-12 04:27:49.329+06	2024-04-12 04:31:29.473+06	7/8/2002	Ikorodu	101242	Lagos
10	00049a81-94a9-4b25-92ed-62d017f3b67a	cristinne.nicor@gmail.com	Female	2024-04-01 16:28:09.574+06	2024-04-01 16:31:47.387+06	8/11/1988	Antipolo	1870	Rizal
11	0005243a-868e-41c4-8d53-3a61649236...	drkhanattaullah@gmail.com	Male	2024-12-23 00:19:56.121+06	2024-12-23 00:28:26.953+06	3/15/1998	Swat	19200	KP
12	00064ab2-0e32-4b50-8d8c-5c6b0d0993...	parthrumrawala@gmail.com	[null]	2025-01-18 10:09:42.03+06	2025-02-09 00:00:40.189+06	[null]	[null]	[null]	[null]
13	000687a6-b0e3-4f21-9df8-98ad0db2ff7e	harshukaurbal@gmail.com	[null]	2024-11-12 00:30:03.689+06	2024-11-12 00:32:23.068+06	[null]	[null]	[null]	[null]
14	000693c5-fafb-4d03-867d-ac6c825db75a	abhijeetsharma8051@gmail.com	[null]	2024-12-11 15:37:39.922+06	2024-12-16 12:20:39.469+06	[null]	[null]	[null]	[null]
15	00075a4d-e755-4ea6-9ef2-e4119495ad49	omolara.onuwaabiamuwe@gmail.com	Female	2024-05-09 04:39:16.09+06	2024-10-10 00:39:05.963+06	12/12/1988	Abuja	900108	FCT
16	00084381-3917-4d03-9639-0a501aa68c...	anushkakushwah18@gmail.com	Female	2025-01-31 00:41:39.305+06	2025-01-31 00:44:11.633+06	9/1/2007	Barwani	451551	Madhya Pradesh
17	000893c1-93cc-425f-b690-ca830f8183da	sophy.thomas60@gmail.com	Female	2024-06-10 01:45:08.634+06	2024-06-10 01:49:40.805+06	3/15/1989	Al Ain	15258	Abudhabi
18	000926e9-66af-4e40-a788-538e74b9a03c	wajihzain.wz@gmail.com	Male	2025-01-23 14:11:31.341+06	2025-01-23 14:17:06.292+06	7/10/1999	Wah Cantt	47100	Punjab
19	0009b4f7-1824-40d9-8391-a7e64bec692f	ralhandivyansha@gmail.com	Female	2024-06-13 09:22:25.777+06	2024-06-13 09:28:26.014+06	4/2/2003	Noida	201309	Noida
20	0009e8ae-95a0-4b58-b5dd-06b90d14ea4f	poozacutee@gmail.com	Female	2025-01-19 09:11:14.781+06	2025-02-18 05:30:41.553+06	8/30/2000	Dhangadhi	10901	Sudurpaschim
21	000aadea-9982-4a02-9851-64ef077756f8	nevernotyours3@gmail.com	[null]	2025-01-12 00:17:26.106+06	2025-01-12 00:17:50.904+06	[null]	[null]	[null]	[null]

Figure 3.22: Preview of the first few rows of the Cognito Data dataset.

	column_name name	data_type character varying
1	user_id	character varying
2	email	character varying
3	gender	character varying
4	user_create_date	timestamp with time zone
5	user_last_modified_date	timestamp with time zone
6	birthdate	character varying
7	city	character varying
8	zip	character varying
9	state	character varying

Figure 3.23: Column names and data types of the Cognito Data dataset.

3.6.2 Gender Distribution

The distribution of gender was analyzed, revealing categories including Male, Female, Other, "Don’t want to specify," and null values. A pie chart visualizes this distribution based on record counts (see Figure 3.24).

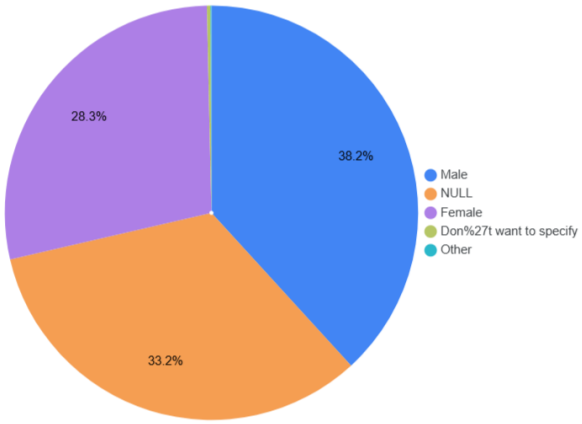

Figure 3.24: Pie chart of gender distribution in Cognito Data.

3.6.3 Distribution of Cities

The top 10 cities by user count were identified, with NULL values being the most frequent, followed by cities like Lagos, Nairobi, and Accra. A bar chart illustrates this distribution (see Figure 3.25).

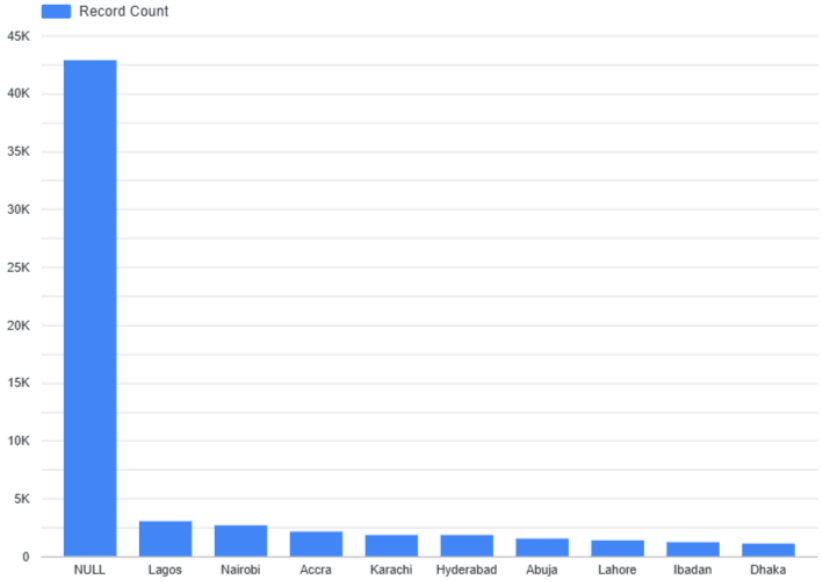

Figure 3.25: Bar chart of the top 10 cities by user count in Cognito Data.

3.6.4 Identification of Missing Values

Missing value counts were calculated for each column, highlighting significant absences in fields such as gender, city, zip, and state. A bar chart visualizes the proportion of missing values across columns (see Figure 3.26).

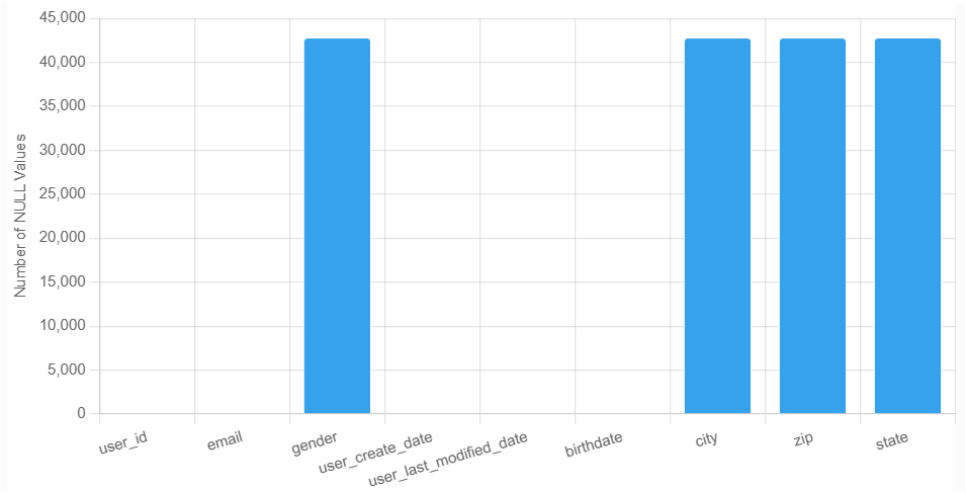

Figure 3.26: Bar chart of missing value proportions in Cognito Data.

3.6.5 Duplicates

Duplicate entries were checked across all columns, confirming no duplicates exist in the dataset, ensuring data uniqueness.

Chapter 4

Data Cleaning Strategy

4.1 User Data

The User Data comprises learner-level data, including unique identifiers, country of origin, education level, institution name, and field of study. A substantial portion of the dataset contains missing values, particularly in the degree, institution, and major columns, with over 40% of the entries lacking one or more of these fields. The learner_id column is clean, consistently formatted, and suitable for uniquely identifying records. However, categorical fields such as degree and major exhibit inconsistencies in naming conventions and capitalization, which require normalization for meaningful analysis. Additionally, several entries contain vague or varied representations of the same institution or field of study, necessitating standardization or grouping strategies. To ensure readiness for downstream analysis, we recommend treating missing values with appropriate labels (e.g., "Unknown"), harmonizing textual data across categories, and optionally deriving new fields such as broader academic disciplines or regional groupings.

4.2 Opportunity Data

The Opportunity Data dataset captures descriptive and logistical information about learning opportunities, including their type, level, provider, location, and access links. While structurally complete, the dataset contains significant inconsistencies in categorical values, date formats, duration expressions, and free-text fields like eligibility. Columns such as type, level, and provider require normalization to unify their respective categories and enable effective filtering. The duration column, currently presented in inconsistent units and formats, should be parsed into a standardized numerical format (e.g., hours or days) for comparative analysis. The deadline and location fields also include mixed formats and missing values, which can be resolved by setting consistent placeholders such as "No Deadline" or "Online." Text-rich columns like eligibility require keyword extraction or text cleaning for usable insights.

4.3 Cohort Data

The Cohort Data dataset captures interaction records between learners and learning opportunities, tracking their application, enrollment, and completion statuses over time. The dataset includes key relational links to both learners (learner_id) and opportuni-

ties (opportunity_id), making it essential for longitudinal analysis. While the structure is clear, the dataset presents formatting challenges, particularly in date fields such as applied_date, enrolled_date, and completion_date, which require standardization and chronological validation. Additionally, the status field contains inconsistent casing and possible typographical variants, which must be unified for accurate status-based filtering. Missing values in enrolled_date and completion_date likely reflect real-world scenarios (e.g., dropped out, ongoing), but should be handled with clearly defined placeholders.

4.4 Marketing Data

The Marketing Data dataset can be cleaned by first resolving any initial parsing issues where all data might be loaded into a single column, then addressing missing values in the campaign_name column by filling them with 'Unknown' and imputing numerical columns like outbound_clicks, landing_page_views, and cpc_cost_per_link_click with their respective medians. Redundant columns such as delivery_level and reporting_starts will be removed due to their constant values. The categorical columns ad_account_name, delivery_status, and result_type will be inspected for consistency; for future steps, it will be recommended to review numerical data for outliers using visualizations, re-verify the uniqueness and consistency of categorical values, and optionally create basic features like conversion rates if relevant for further analysis.

4.5 Learner Opportunity Data

The Learner Opportunity Data dataset captures post-engagement feedback from learners regarding their experiences with various learning opportunities. It forms a key bridge between learner and opportunity datasets, enabling sentiment analysis and outcome evaluation. However, the dataset requires significant preprocessing to ensure reliability. The learner_feedback column contains free-form text that varies in length and structure; it can be cleaned for spelling, punctuation, and analyzed for sentiment or topic categorization if deeper insights are desired. The satisfaction field mixes textual ratings with potential numeric values, necessitating normalization to a unified scale (e.g., 1-5 or "Low" to "High"). Missing values in both columns should be handled gracefully, labeled as "No feedback" or "Not Rated" to maintain analytical consistency.

4.6 Cognito Data

The UserCreateDate and UserLastModifiedDate columns will be converted to datetime objects for temporal analysis, and missing values in gender, birthdate, city, zip, and state will be filled with 'Unknown'. Duplicate email addresses will be identified and addressed, and gender categories like 'Don%27t want to specify' will be standardized to 'Other'. For next steps, decisions will be made on handling duplicate emails, birthdate column will be further refined if age calculations are crucial, and the zip column's data type will be considered for numerical operations or geographical analysis if needed.

Chapter 5

Conclusion

This report, prepared by Team 26, consolidates the data visualization and analysis efforts, fulfilling the deliverables . Completed analyses for User Data and Cognito Data provide insights into user demographics and data quality issues, respectively. As analyses for Opportunity Data, Cohort Data, Marketing Data, and Learner Opportunity Data are received, they will be integrated to offer a comprehensive overview of the learning platform's performance and user engagement. The findings lay the groundwork for Week 2, focusing on transforming cleaned datasets for advanced analysis and visualization.