Internship Report

Data Visualization and Analysis

Week- 3

Date: 23-06-2025

Team Number: 26

Team Members:

1. Khushi Khati (khushikhati11@gmail.com)
2. Rohit Emmanuel (kingrohit2439@gmail.com)
3. Shomitra Dey (soumitradev532@gmail.com)

Table Of Contents:

Abstract

This report, prepared by Team 26, details the Week 3 progress of our internship project focused on designing and implementing a robust data integration process. Our primary objective for the week was to construct a unified and consistent core dataset from the cleaned datasets. To achieve this, we began by creating a detailed mapping table in Excel, which served as a blueprint for understanding the structure and relationships across datasets, identifying key columns like learner_id, enrollment_id, opportunity_id, and documenting them. Notably, we also performed Exploratory Data Analysis (EDA) on the marketing dataset and subsequently cleaned it, preparing it for potential future integration or standalone analysis, despite its initial lack of relational keys to the other datasets. A significant learning experience involved an experiment where we built our integrated dataset twice: once directly from raw datasets and once from pre-cleaned datasets. This exercise underscored the value of preprocessing, as the cleaned version yielded improved column consistency, fewer NULLs, and enhanced usability, reinforcing our iterative learning approach. Furthermore, we conceptualized the future use of this integrated dataset by creating a prototype wireframe for potential dashboards. This involved grouping columns from our integrated dataset into logical categories to ensure ease of analysis and visualization for future analytical work. Throughout this process, we applied structured SQL queries to detect data issues, standardize formats, and validate the integrity of the final integrated dataset. The completed unified dataset now serves as a reliable foundation for future analytical work and dashboards, and this report highlights both our technical decisions and our iterative learning approach toward building scalable and dependable data pipelines.

CHAPTER 1

Introduction

This report, prepared by Team 26, outlines the comprehensive progress made in Week 3 of our internship project, focusing on the design and implementation of a robust data integration process. Our overarching objective for the week was to construct a unified and consistent core dataset from several disparate, yet cleaned, datasets provided. This unified dataset is now a reliable foundation for future analytical work and dashboard development, embodying both our technical decisions and an iterative learning approach.

To achieve our primary objective, we initiated the process by creating a detailed mapping table in Excel. This table served as a foundational blueprint, meticulously documenting the structure and intricate relationships across the various input datasets. It was instrumental in identifying and categorizing key columns, such as `learner_id`, `enrollment_id`, and `opportunity_id`, which are crucial for establishing logical connections between different data entities.

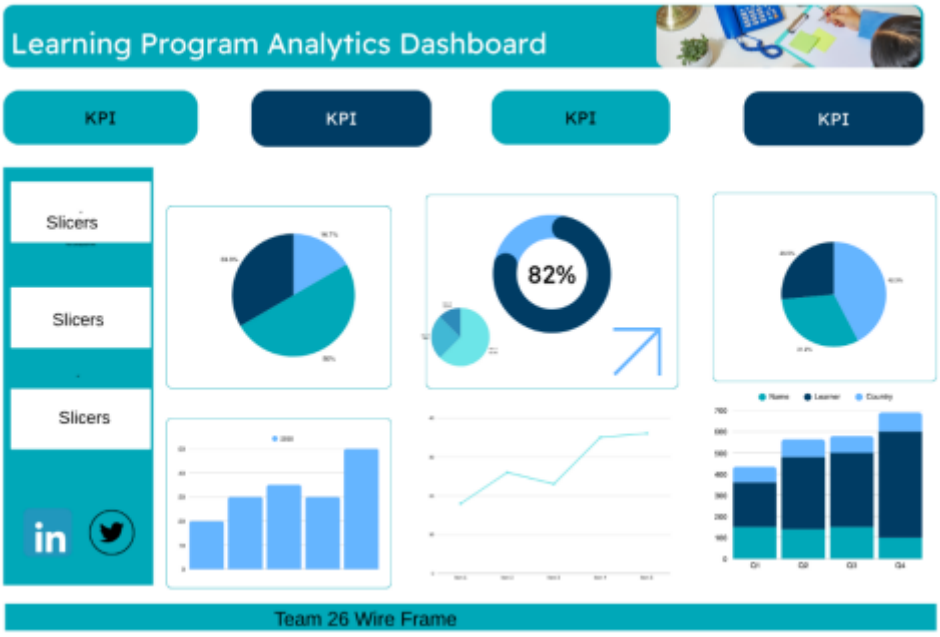
A significant phase of our work involved performing Exploratory Data Analysis (EDA) on the marketing dataset. Following the EDA, this dataset underwent a thorough cleaning process, preparing it for potential future integration or standalone analysis. It's important to note that while cleaned, this marketing dataset initially lacked direct relational keys to the other core datasets, suggesting its potential use for distinct analytical perspectives or future, more complex integrations.

A pivotal learning experience emerged from an experiment where we built our integrated dataset twice. The first attempt involved direct integration from raw datasets, while the second utilized pre-cleaned datasets. This exercise profoundly underscored the immense value of preprocessing data. The version constructed from pre-cleaned data consistently yielded superior results, demonstrating improved column consistency, significantly fewer NULL values, and greatly enhanced usability. This practical demonstration reinforced our team's commitment to an iterative learning approach, emphasizing the importance of clean input data for robust integration.

Furthermore, we applied structured SQL queries throughout the process. These queries were not only used to detect and rectify various data quality issues but also to standardize data formats and validate the overall integrity of the final integrated dataset. This rigorous validation ensures that the unified dataset is dependable and accurate for any subsequent analytical tasks.

CHAPTER 2

Wireframe



2.1 Creation Of The Wireframe

To visualize data insights about users, opportunities, and program engagement based on the multiple CSV datasets (user, cohort, etc.).

- **Slicers (Left Section):** We added slicers to let users interact with the dashboard and filter insights based on key variables. Slicers make the dashboard dynamic and customizable, helping stakeholders drill down to what matters most to them.

- **KPIs (Top Row):** These 4 KPI boxes at the top summarize the most important metrics at a glance. These KPIs act as a quick status check; they help decision-makers instantly grasp the program's performance without digging into charts.

- **Charts (Middle & Bottom Section):** We included a mix of Pie Charts, Donut Charts, Bar Charts, and Line Charts. These visualizations let stakeholders spot patterns, outliers, and growth opportunities easily across various datasets.

- **Social Media Icons (Bottom Left):** LinkedIn and Twitter icons represent external outreach and visibility. You likely added them to indicate that this dashboard might be shared or connected to social performance metrics. It shows a connection between your data and public-facing results, especially useful for marketing or reporting to sponsors/partners.

- **Team Tag:** Team 26 Wire Frame at the bottom gives credit to your group and signals collaboration. It gives stakeholders context on ownership, accountability, and the ability to reference the right team.

2.2 Summary

This dashboard wireframe is structured for clarity, interactivity, and storytelling. It reflects a professional, data-driven approach to tracking and communicating the impact of a learning program from recruitment to engagement and completion.

CHAPTER 3

Master Table

3.1 Revised Master Table

In our master table, the following columns were dropped: status, lo_learner_id, cognito_user_id, opportunity_id, opportunity_code, cohort_code, tracking_questions, size
The datasets outline a comprehensive system for managing learners, the opportunities they engage with, and their grouping into cohorts, with Cognito handling user profiles details. The learnerOpportunity file serves as the central link, connecting individual learners to specific opportunities and their respective cohorts through fields like learner_id, assigned_cohort, and enrollment_id.

The Cognito dataset, with its user_id, likely complements the Learner dataset by providing additional user profile data, assuming user_id corresponds to learner_id.

This is the link to the revised Master Table

3.2 Grouping Of Columns From The Master Table:

The following link outlines five specialized sub-tables derived from a main master table, each designed to provide specific analytical insights for dashboards. These sub-tables include Regional Distribution for geographic analysis, Demographics for characteristic analysis like gender and age, Academic Background for educational profiles, Opportunity Details for tracking program enrollments and application patterns, and User Account Information for account activity and data quality insights. Most sub-tables are designed to hold unique `learner_id` entries to avoid data redundancy, except for Opportunity Details, which captures multiple enrollments per learner.

This is the link to the sub-tables

CHAPTER 4

Mapping Table

The datasets outline a comprehensive system for managing learners, the opportunities they engage with, and their grouping into cohorts, with Cognito handling user profile details. The learnerOpportunity file serves as the central link, connecting individual learners to specific opportunities and their respective cohorts through fields like learner_id, assigned_cohort, and enrollment_id.

The Cognito dataset, with its user_id, likely complements Lthe earner dataset by providing additional user profile data, assuming user_id corresponds to learner_id.

This foundational user identity, typically represented by a user_id, is then linked to the learner_id in the Learner dataset, which further enriches the profile with specific academic or professional background information relevant to their role as a learner. As learners seek engagement, they interact with the Opportunity dataset, a dataset detailing various available programs, courses, or roles.

The critical connection between a learner and an opportunity is meticulously recorded in the LearnerOpportunity dataset. This file serves as an enrollment or application registry, documenting which learner_id is associated with which opportunity, and crucially, assigns them to a specific cohort_code as defined in the Cohort dataset. The Cohort dataset then provides the contextual grouping for learners, indicating the timeframe (start_date, end_date) within which they participate in an opportunity.

In essence, the flow moves from user creation (Cognito) to learner profiling, then to opportunity discovery, enrollment tracking (LearnerOpportunity), and finally, grouping within specific cohorts. This interconnectedness allows for comprehensive tracking of a learner's journey, from initial sign-up to their progression through various educational or professional opportunities.

This is the link to the Mapping Table

CHAPTER 5

Marketing Dataset

This chapter involves the summary, cleaning process, EDA, issues that arose while performing EDA on the Marketing dataset, recommended solutions to solve the issue, and the reasons why this dataset was not included in the master table.

5.1 Summary of the Marketing Campaign Dataset

This dataset contains performance data for digital marketing campaigns managed by the ad account SLU. The campaigns appear to focus on job postings and brand awareness (e.g., internships, website leads, UGC videos).

5.1.1 Key Columns:

- **Campaign name**: Describes the campaign's focus (e.g., "Internship", "Brand Awareness").
- **Delivery status**: Whether the campaign is *active*, *completed*, or *inactive*.
- **Reach**: Number of people who saw the campaign.
- **Outbound clicks**: Clicks that led users to external landing pages.
- **Landing page views**: Users who visited and viewed the landing page.
- **Results**: Campaign-specific outcome (e.g., leads, website applications).
- **Cost per result**: How much was spent per desired action.
- **Amount spent (AED)**: Total ad spend in AED currency.
- **CPC (Cost per click)**: Cost for each link click.
- **Reporting starts**: Start date of campaign tracking

5.1.2 Key insights:

- Some campaigns have very high reach (e.g., 18 million+ impressions).
- Performance metrics vary significantly across campaigns (e.g., CPC from 0.04 to 0.40 AED).
- Results types include leads, applications, and reach-based metrics.

5.2 Why is the Marketing Campaign not included in the Master Table?

This dataset was not included in the master table because it represents performance data from paid digital marketing campaigns, which is structurally and contextually different from the organic or platform-specific datasets used in the main analysis. While the master table likely focused on user engagement, content performance, or audience behavior across platforms like Instagram or Facebook, this dataset centers on ad-level metrics such as cost-per-click (CPC), ad spend, and campaign reach—attributes that are not directly linked by shared identifiers like post IDs, user IDs, or timestamps.

However, this dataset is still useful as it offers insights into paid marketing efficiency, showing how much was spent and what results were achieved across campaigns. It helps evaluate ROI, cost-effectiveness, and audience response to ad-driven content, which could complement—but not directly merge with—the organic datasets in the master table. Including it in the master table without proper relational keys or shared variables would introduce inconsistency and limit meaningful cross-comparison.

5.3 Cleaning the Marketing

The SQL queries used for cleaning the marketing campaign dataset can be accessed via the following link:
SQL Queries
These queries handled tasks such as:

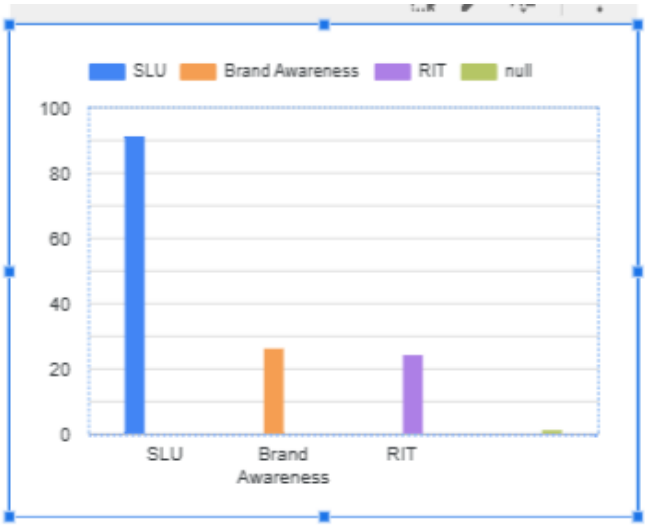
- Removing undefined/null entries
- Converting text fields to proper case
- Standardizing metrics (e.g., CPC, cost per result)
- Filtering out incomplete or irrelevant campaign records

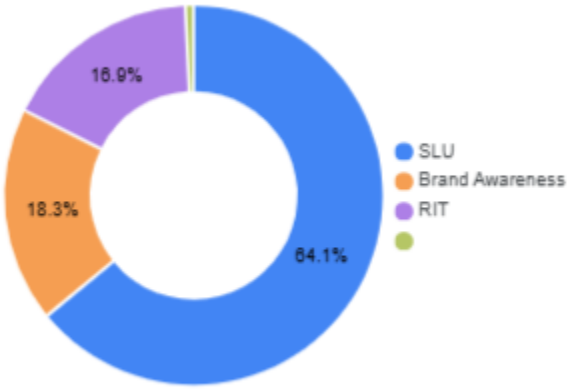
This ensures the dataset is clean, consistent, and ready for meaningful analysis

5.4 Exploratory Data Analysis

1.

Column Name: Ad Account Name
Chart Type: Bar Chart
Description: Displays the count of marketing entries per ad account. Bar chart accurately shows all accounts, including nulls.
Note: In pie charts, NULL values are incorrectly shown as `'undefined'` due to Looker Studio's handling of nulls in TEXT columns.



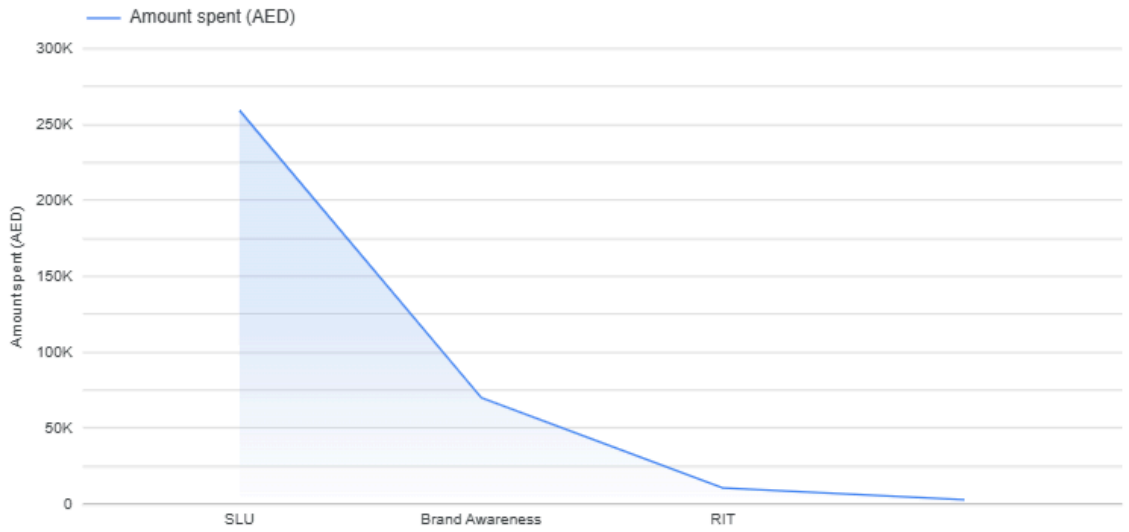


2.

Column Name: Amount Spent

Chart Type: Line Chart

Description: Shows spending trends over time for each campaign. Useful for identifying peaks and dips in expenditure patterns. Enables time-based comparison of budget allocation across campaigns.

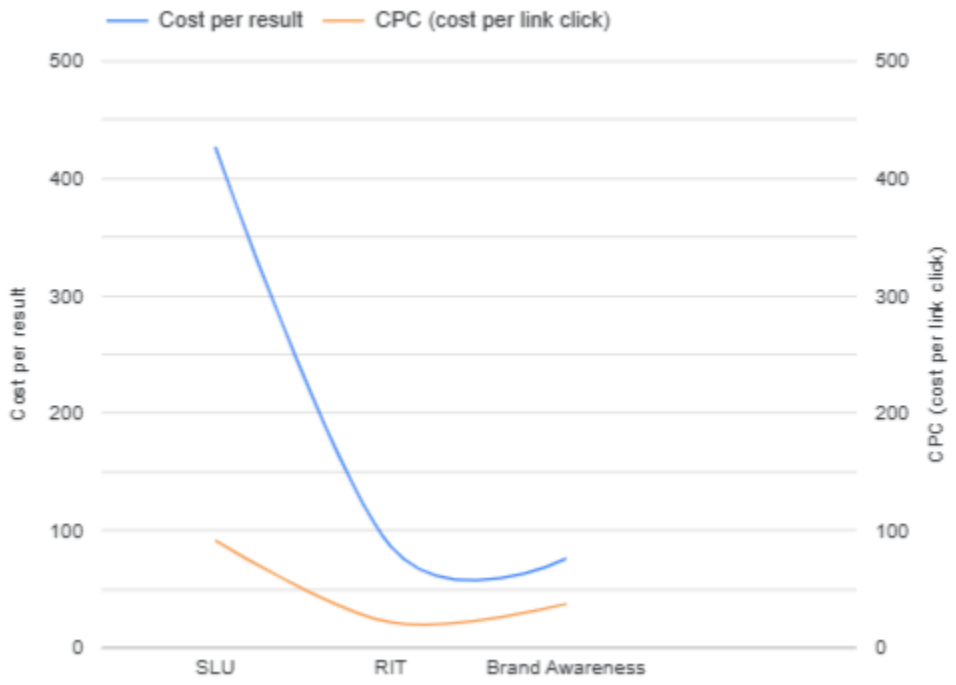


3.

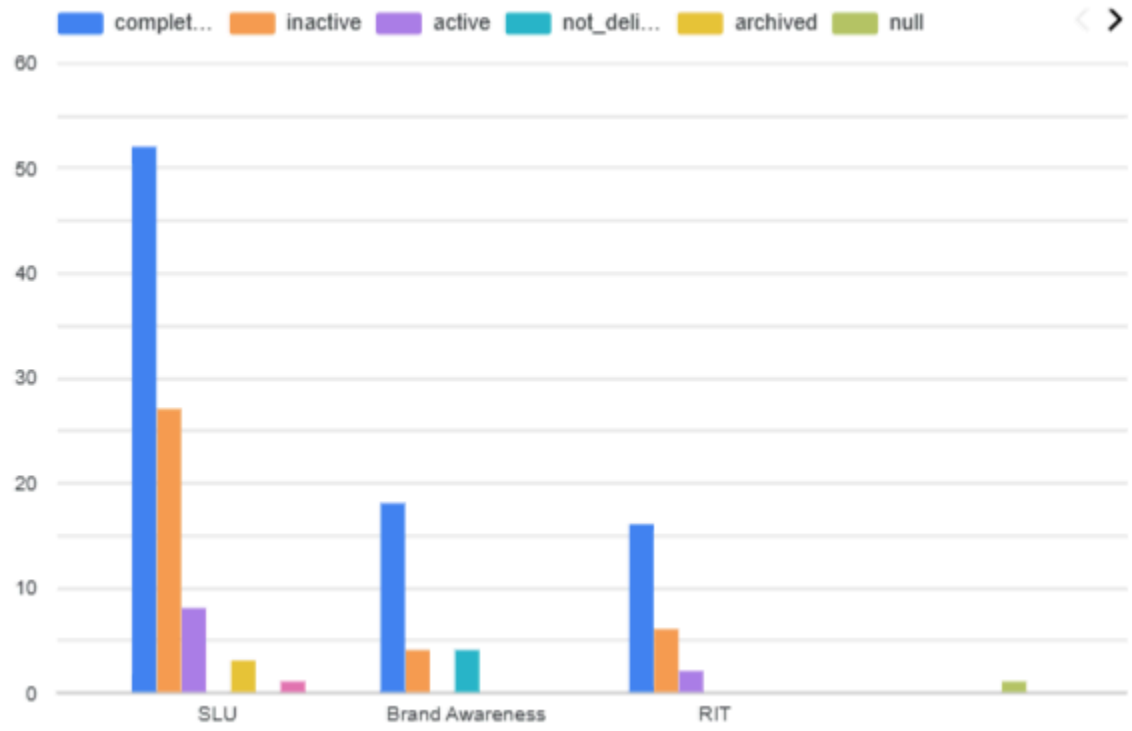
Column Name: Cost per Result, Cost per Click

Chart Type: Line Chart

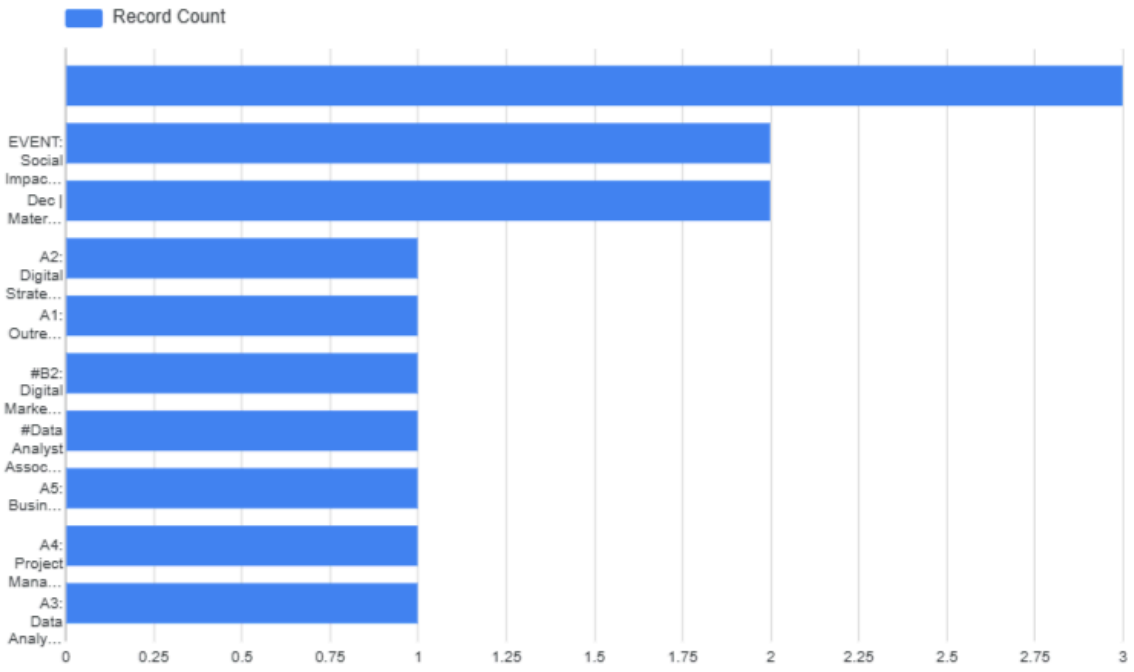
Description: Illustrates cost efficiency trends across campaigns over time. Helps compare the effectiveness of ad spend by tracking fluctuations in cost per engagement and cost per action.



4.
Column Name: Delivery
Chart Type: Bar Chart
Description: Shows the current delivery status (e.g., Active, Completed, Inactive) for each ad account. Useful for monitoring campaign lifecycle and identifying which accounts have ongoing or paused activities.

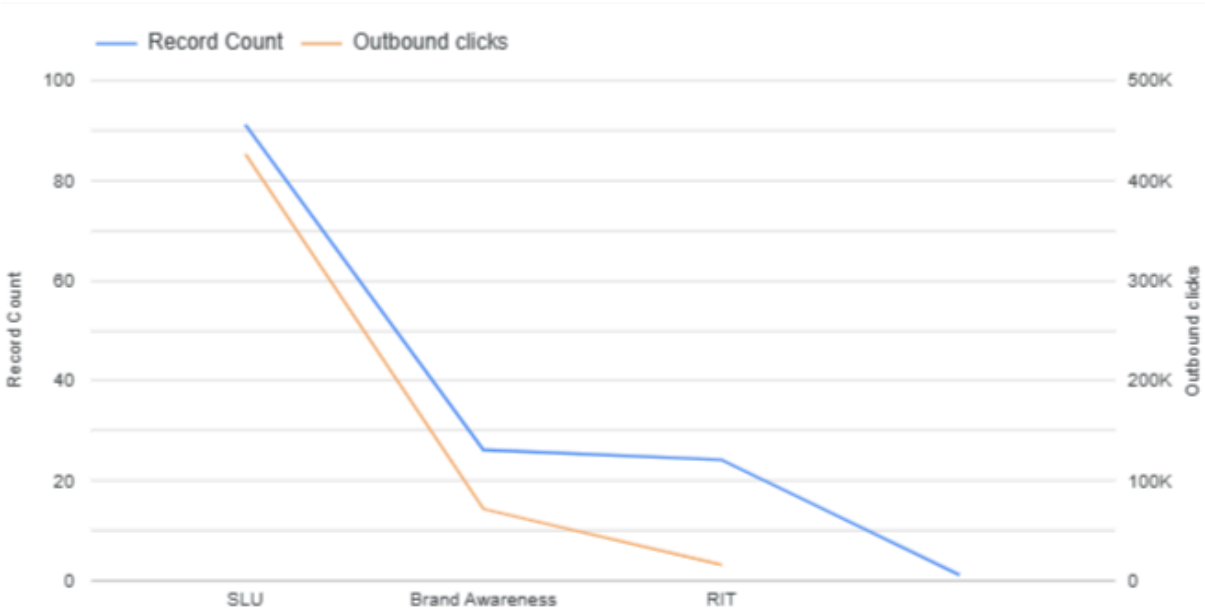


5.
Column Name: Campaign Name
Chart Type: Bar Chart
Description: Displays the frequency of records per campaign. Helps identify the most active or frequently used campaigns across the dataset for strategic assessment.



6.

Column Name: Outbound Clicks, Record Count
Chart Type: Line Chart
Description: Tracks user engagement (clicks) alongside the number of records over time. Useful for identifying patterns in audience interaction and data volume trends across campaigns.



7.

Column Name: Ad Account Name, Record Count, Amount Spent, Cost per Result, Cost per Click, Landing Page

Chart Type: Summary Table

Description: Provides a consolidated view of key performance metrics by account. Enables comparison of spend, efficiency, and landing page usage across different ad accounts in a single, actionable format.

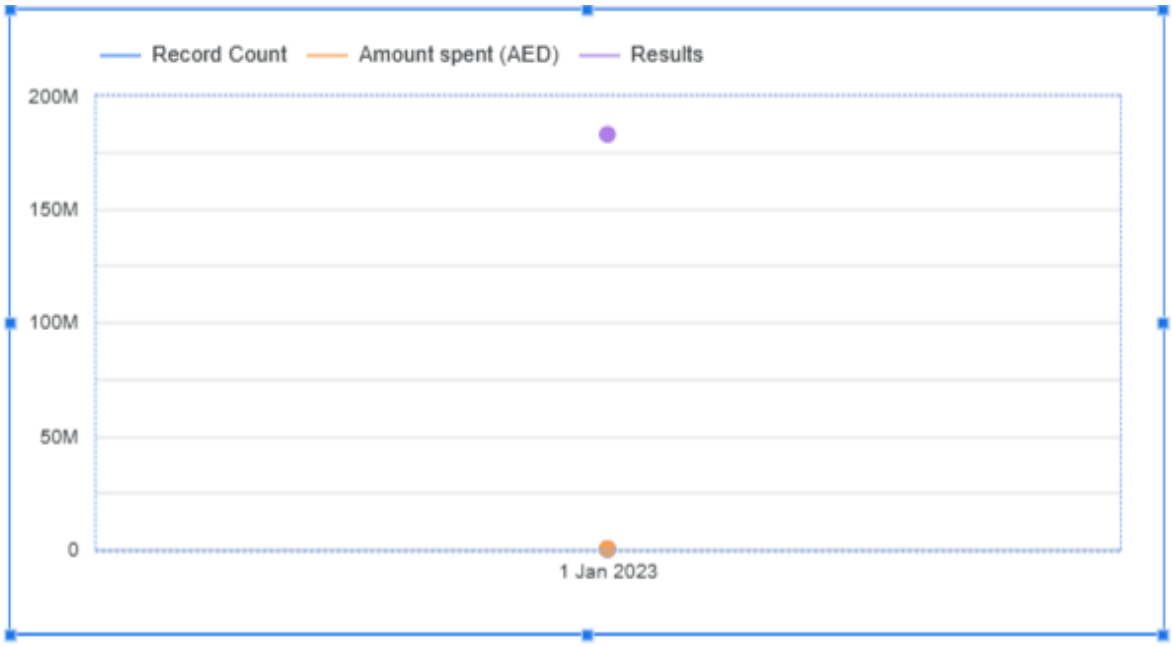
	Ad Account Na...	Record ...	Amount sp...	Cost per result	CPC (cost per link click)	Landing pag...
1.	SLU	91	258,760.88	426.28	90.63	255,215
2.	Brand Awareness	26	69,507.92	75.11	36.52	26,961
3.	RIT	24	10,124.91	85.73	20.78	11,666
4.	null	1	2,436.47	null	null	null

8.

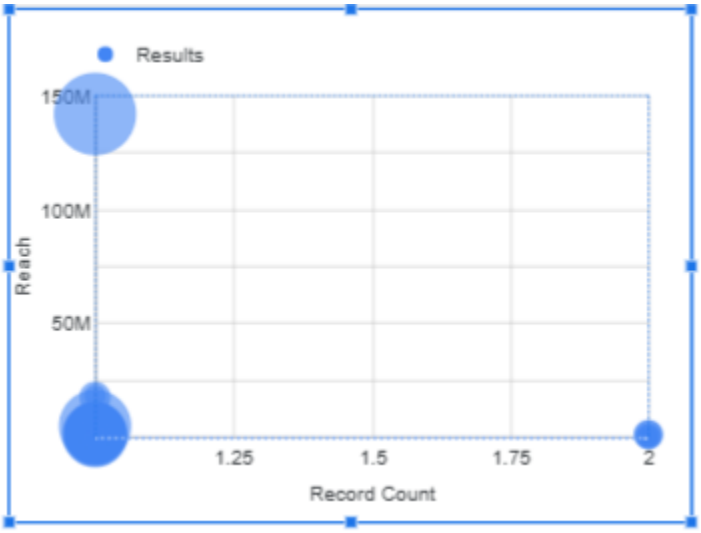
Column Name: Amount Spent, Results, Reporting Starts

Chart Type: Time Series Chart

Description: Visualizes the relationship between spending and campaign results over time. Helps assess the performance impact of budget allocation across different reporting periods.



9.
Column Name: Amount Spent, Reach
Chart Type: Scatter Chart
Description: Plots spending against audience reach to identify correlation patterns. Useful for evaluating the efficiency of the budget in driving campaign visibility.



5.5 Issues Faced

During exploratory data analysis (EDA), we encountered a specific issue related to TEXT-type columns. As part of the data cleaning process, we standardized missing or invalid entries by checking for values such as `'undefined'`, `'NaN'`, `'null'`, and empty strings. These were all converted to SQL-compliant `NULL` values. To ensure consistency, we also applied trimming of whitespace and transformed all text to lowercase before performing these checks, thereby eliminating any variation in representation.

However, during visualization in Looker Studio, we observed that pie charts incorrectly labeled `NULL` values as `'undefined'`, despite our preprocessing. In contrast, bar charts and other visualizations correctly recognized and displayed `NULL` values without issue.

Upon further investigation, we discovered that Looker Studio sometimes interprets `NULL` values in TEXT columns as `'undefined'` during rendering, particularly in charts like pie charts which rely on grouping text categories. This appears to be a known behavior due to how Looker internally converts `NULL` values to strings in some chart types.

5.6 Recommended Workaround

- Prefer using bar charts or scorecards when dealing with NULL values in text columns, as they offer more accurate rendering.
- Alternatively, consider applying a CASE WHEN or COALESCE() transformation in your SQL or Looker-derived table to explicitly label NULL values as `'Missing'` or `'No Data'` prior to visualization.

Example:

sqlCopyEditCOALESCE(column_name, 'No Data') AS cleaned_column

This ensures consistent and controlled display of missing values across all chart types.

CHAPTER 6

Conclusion

Team 26's Week 3 efforts successfully culminated in the creation of a reliable, unified dataset. This process not only involved meticulous data integration and cleaning but also a deep dive into the business logic connecting each data source. The lessons learned, particularly the value of preprocessing, and the proactive planning for future analytical tools like dashboards, underscore our commitment to building scalable and dependable data pipelines for effective business intelligence.

Looking ahead, we conceptualized the future use of this integrated dataset by creating a prototype wireframe for potential dashboards. This foresight involved thoughtfully grouping columns from our unified dataset into logical categories. This strategic organization ensures that future analytical work and data visualization will be intuitive, enabling stakeholders to easily extract insights regarding learner engagement, opportunity performance, and cohort progress.