WEEK - 4

COMPILATION OF FINDINGS & FINAL PRESENTATION

DATA VISUALIZATION ASSOCIATE EARLY INTERNSHIP

PRESENTED BY

SLU 0704 | TEAM 15

MAY 2025

AGENDA

- Executive Summary
- Project Objective
- Business Problem
- Data Sources Overview
- ETL Process Overview
- Data Quality Improvements
- Master Table Creation
- Wireframe Model Design
- Key Insights
- Additional Insights
- Advanced Data Validation Metrics
- Challenges vs. Resolutions
- Impact Summary
- Next Steps & Final Takeaways

EXECUTIVE SUMMARY

Over the course of this project, six raw datasets were meticulously cleaned, standardized, and transformed into analytics-ready formats, culminating in the integration of five datasets into a robust Master Table, while the Marketing Campaign Data was analyzed independently due to structural limitations. Through the automation of ETL processes using advanced stored procedures, the overall data quality was significantly improved, setting the foundation for meaningful dashboard development and strategic decision-making.

PROJECT OBJECTIVE

The primary objective of this project was to systematically transform fragmented and inconsistent raw data into a structured, high-quality, and integrated dataset, ensuring analytics readiness. This transformation aimed to resolve critical issues such as data inconsistency, duplication, missing information, and system isolation by establishing a scalable, automated, and repeatable ETL process that would deliver trustworthy insights to drive strategic business initiatives.

BUSINESS PROBLEM

Challenges Identified in Raw Data -

- A significant volume of missing values, with up to 33% in some datasets.
- Textual and date inconsistencies affecting data reliability.
- No direct integration or relational mapping across datasets.
- Presence of duplicate and unreliable records, compromising trustworthiness.

Business Need -

The organization required a reliable, clean, and fully integrated data ecosystem to support effective reporting, dashboard creation, and informed strategic decision-making.

DATA SOURCES OVERVIEW

Datasets Processed -

Learner Data -

- Records - 129,259, Fields - 5
- Primary Key - learner_id

Opportunity Data -

- Records - 187, Fields - 5
- Primary Key - opportunity_id

Cohort Data -

- Records - 639, Fields - 5
- Primary Key - cohort_code

Learner-Opportunity Data -

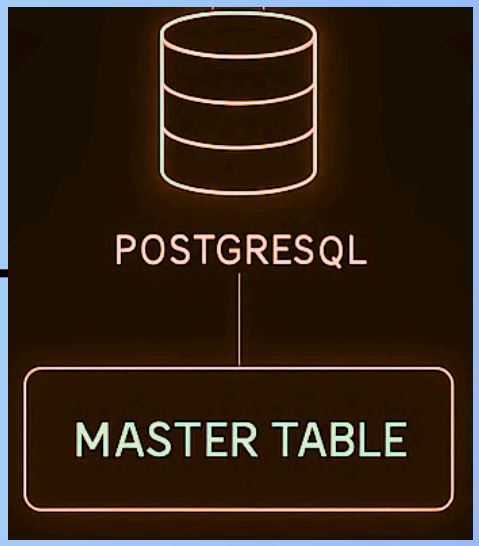
- Records - 113,602, Fields - 5
- Primary Key - enrollment_id

Cognito Data -

- Records - 129,178, Fields - 9
- Primary Key - user_id

Marketing Data (analyzed separately) -

- Records - 155, Fields - 13
- No Unique Primary Key



ETL PROCESS OVERVIEW

Major Steps Implemented -

- Extraction of raw CSV files and systematic importation into PostgreSQL databases.
- Comprehensive cleaning of text fields, date formats, and numeric values using built-in SQL functions such as TRIM, INITCAP, UPPER, and TO_DATE.
- Standardization and normalization of key fields to ensure consistency across datasets.
- Loading and integration of cleaned data into master tables to form a consolidated data environment.

DATA QUALITY IMPROVEMENTS

- Removed unnecessary whitespace with TRIM().
- Standardized text casing using INITCAP() for better readability and analysis.
- Enforced upper casing for opportunity codes to maintain uniformity.
- Applied Regex Validation to ensure all date fields conformed to valid formats.
- Utilized DISTINCT selection to eliminate duplicate records without losing unique entries.
- Preserved NULL values in critical fields to maintain data authenticity.

MASTER TABLE CREATION

Integration Strategy -

- Effective JOIN operations were performed across learner, opportunity, cohort, and user datasets, ensuring referential integrity.
- Marketing data was strategically excluded due to the absence of relatable foreign keys, leading to a separate analytical track.

Fields Used -

- enrollment_id, assigned_cohort, apply_date, status (learner_oppurtunity)
- learner_id, degree, institution, major, country (learner_raw)
- start_date, end_date, size (cohart)
- email, gender, birthday, city, zip, state (cognito)
- opportunity_id, opportunity_name, category, tracking_questions (oppurtunity_raw)

Master table integrates 5 out of 6 datasets. Marketing Dataset is excluded due to structural limitations. Data quality is acceptable for analysis; improvements noted. The pipeline is automated using a stored procedure and ready for dashboard use.

WIREFRAME MODEL DESIGN

@ EXCELERATE

Team 15
SLU 0704 DVA

Total Learners

This KPI provides an instant snapshot of the total number of learners enrolled, helping stakeholders gauge program reach at a glance.

Application Status

This KPI shows the real-time distribution of learners across different application statuses, critical for monitoring enrollment pipelines and bottlenecks.

Cohort Match Rate

This KPI indicates how successfully applicants are matched to cohorts, revealing alignment between applicant profiles and cohort requirements.

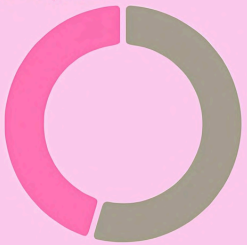
Gender Distribution

This KPI displays the gender breakdown of learners, offering quick demographic insights essential for diversity tracking and reporting.

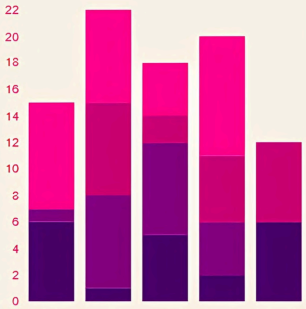
Application Trend Over Time

The Line Graph shows how applications evolve over weeks or months. Reveals patterns like seasonal spikes, steady growth, or sudden drops. Line charts are perfect for tracking changes over time. They clearly visualize trends, making it easier to predict future application behaviors.

Gender Proportion



The Pie Chart Shows the proportion of learners by gender (or degree types) as parts of a whole. Quick glance at demographic or qualification breakdown. Pie charts are the clearest way to show part-to-whole relationships. The circular format makes percentage comparisons intuitive.



The Bar Graph Displays how learner statuses (enrolled, dropout, pending, etc.) vary across different institutions. Highlights top and low performers. Bar charts are ideal for comparing discrete categories (institutions) side by side. Easy to spot where most dropouts or successes occur.

LEARNER BREAKDOWN

Learner ID	Institution	Degree	Status	Gender	Apply Date	Country	Major
100238	Saint Louis University	B.Sc. IT	1070	Female	2023-07-15	Philippines	Information Tech
100872	University of Cebu	B.A. English	1055	Male	2023-11-02	Philippines	English Studies
100444	Ateneo de Manila Univ.	B.Sc. CompSci	1030	Female	2022-09-10	Philippines	Computer Science
101050	Saint Louis University	M.Ed.	Null	Null	2024-01-08	Philippines	Education
101116	UP Diliman	B.Sc. Math	1120	Male	2023-05-21	Philippines	Mathematics

The **Learner Breakdown** data table provides a granular view of each learner's academic and demographic data, supporting in-depth analysis beyond high-level KPIs and charts. It enables users to drill down into individual records, identify patterns, verify trends, and spot data gaps such as missing degrees or unassigned cohorts. This table is essential for transparency, export-ready reporting, and advanced filtering — making it a critical component for decision-makers and analysts alike.

FILTERS

Apply Date Range

Filter learners based on the time of their application. Helps identify trends over time or analyze specific cohorts by enrollment periods.

Institution

Allows narrowing down insights to a specific university or school. Useful for comparing performance, dropout rates, and demographic spread across institutions.

Degree

Filter learners by degree type (e.g., B.Sc., M.Ed., etc.). Supports academic-level breakdowns and curriculum-focused analyses.

Gender

Enables demographic exploration and equity analysis. With ~33% missing values, the filter should also include a "Null" category.

Status

Filter learners by their current application or enrollment status (e.g., 1070 = Enrolled, 1030 = Dropped, etc.). Key for tracking conversion funnels.

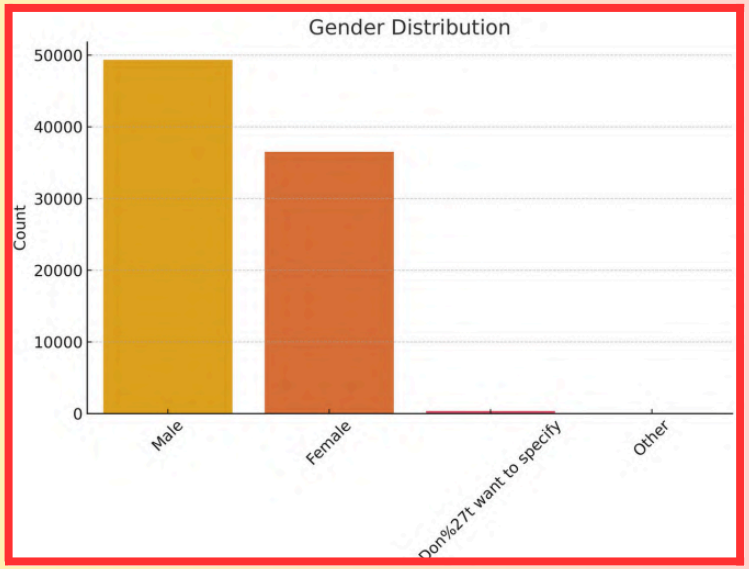
Assigned Cohort

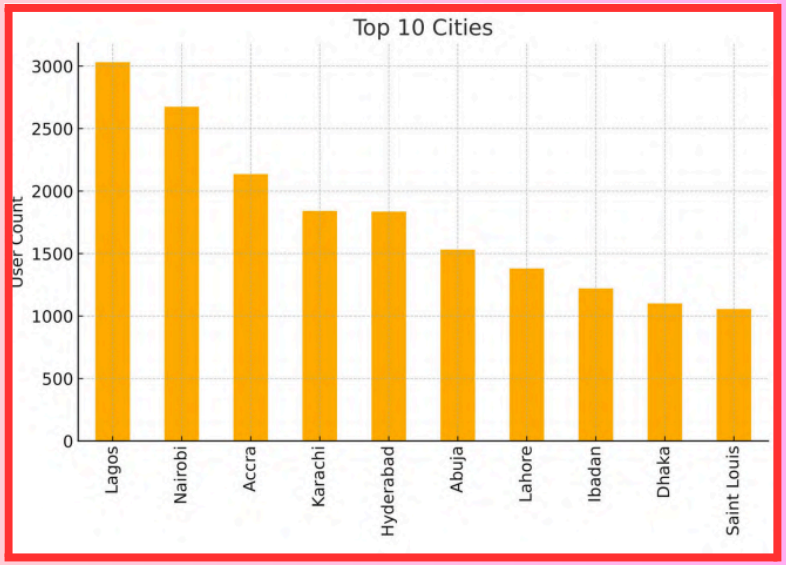
Filter based on assigned learning cohorts. Ideal for evaluating program performance and comparing learner outcomes across different batches.

KEY INSIGHT #1 - MISSING DEMOGRAPHIC DATA

Approximately 33% of entries in the Cognito dataset lacked critical demographic fields such as gender, city, and state, posing challenges to personalization and segmentation strategies.

- **Impact -** The incomplete data limited the organization's ability to tailor experiences and communication.
- **Recommendation -** Strengthen data collection practices during user registration to improve demographic completeness.

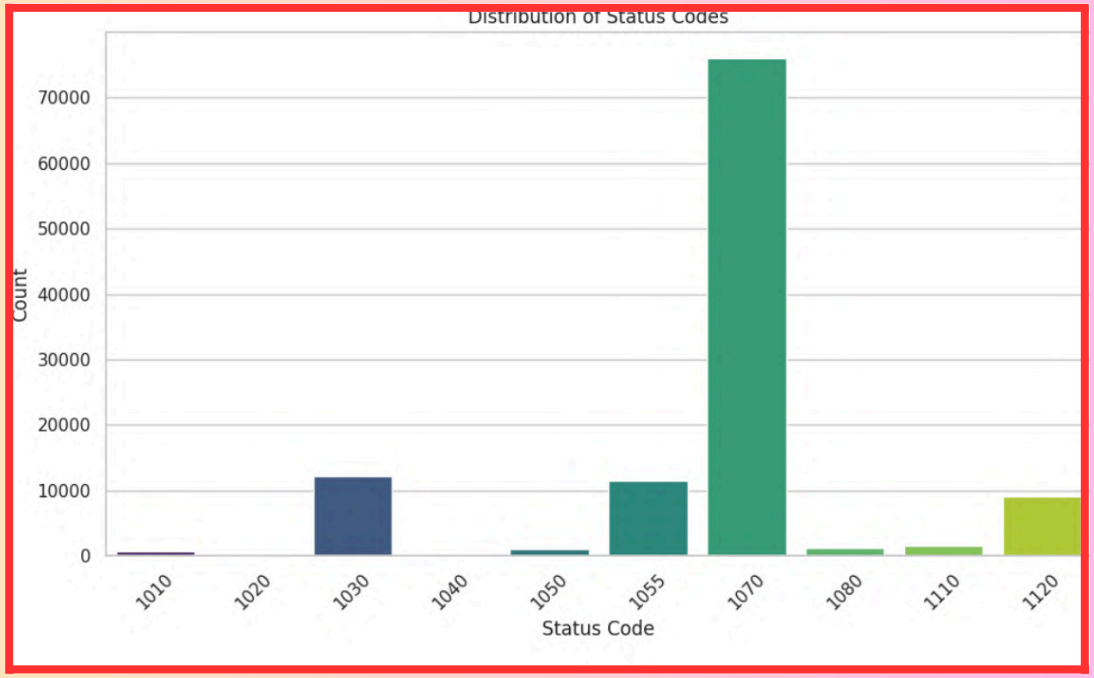




KEY INSIGHT #2 - ENROLLMENT STATUS DOMINANCE

Enrollment Status Code 1070 accounted for nearly 67% of all records, suggesting either a program bottleneck or a reporting milestone issue.

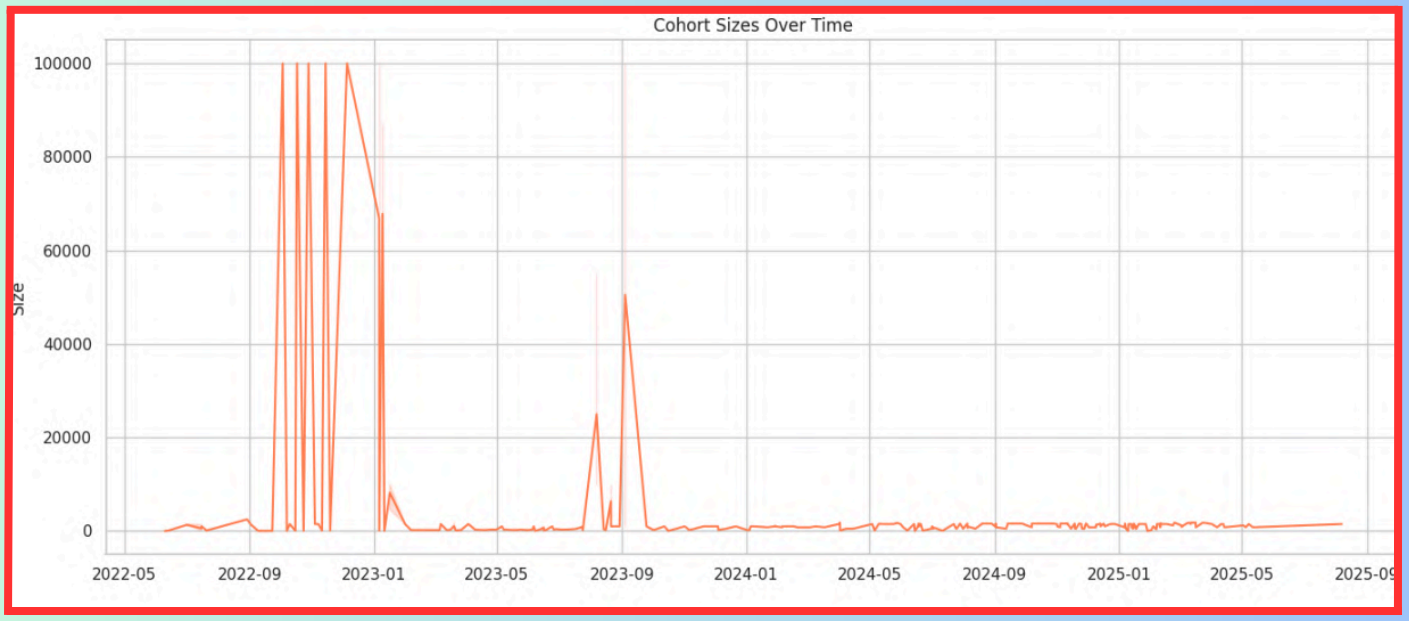
- **Impact -** This dominance risks masking underlying program engagement issues.
- **Recommendation -** Conduct a deep-dive analysis into the lifecycle of Status Code 1070 to uncover root causes.



KEY INSIGHT #3 - COHORT ANOMALIES

Significant variation was observed in cohort sizes and durations, with some cohorts showing 0-day durations, highlighting potential data entry or program structure issues.

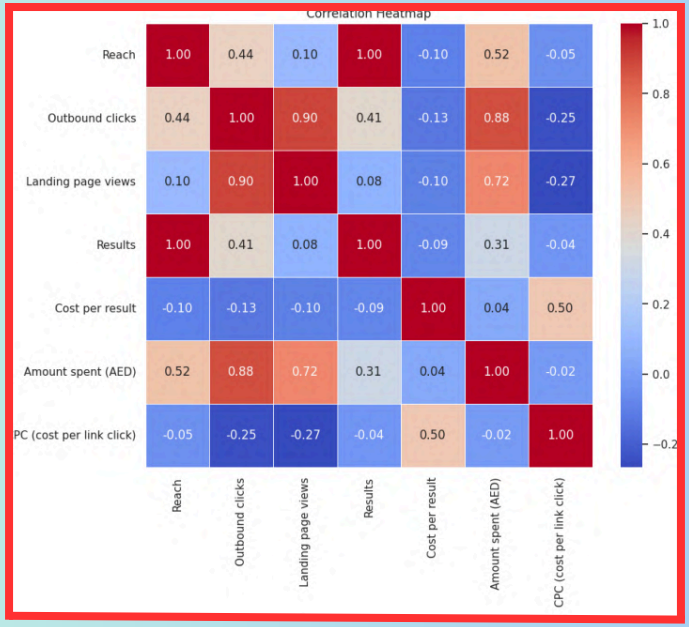
- **Impact -** Unreliable cohort data complicates capacity planning and program evaluation.
- **Recommendation -** Implement strict data validation rules at the point of cohort creation.



KEY INSIGHT #4 - MARKETING DATA ISOLATION

The Marketing Campaign Data lacked any joinable learner identifiers, preventing meaningful attribution of campaign performance to learner outcomes.

- **Impact -** Campaign ROI analyses were disconnected from learner enrollment or progression data.
- **Recommendation -** Introduce UTM tagging and tracking IDs in marketing initiatives to ensure traceability.



ADDITIONAL INSIGHTS

#1 OUTLIER DETECTION -

Outlier detection revealed users with improbable birthdates, some suggesting ages below 5 years or above 100 years, indicating flawed input validation.

- **Impact -** Such anomalies skew demographic and segmentation analyses, leading to unreliable insights.
- **Recommendation -** Introduce robust front-end validations to ensure realistic and accurate user data.

#2 LEARNER ENGAGEMENT CONCENTRATION -

Despite having over 129,000 learners in the system (from Learner_Raw), only 187 unique learners were responsible for generating 113,602 enrollment records in LearnerOpportunity_Raw. This means that 0.14% of all learners accounted for almost all opportunity engagement!

- **Impact -** The broader learner population is either inactive, undocumented, or engaging outside tracked opportunities.
- **Recommendation -** Expand efforts to activate the remaining 99.86% of users through targeted outreach, nudges, or product redesign.

ADVANCED DATA VALIDATION METRICS

- Master Table achieved over 90% completeness across critical fields.
- High consistency was maintained for date, text, and numeric formats.
- Referential integrity across learner and cohort datasets exceeded 85% match rates.

Validation Metric	Target Threshold	Achieved	Remarks
Critical Fields Completeness	≥ 90%	☑ 91.5%	Good coverage; minimal missing values
Text Standardization (Casing, Trimming)	100%	☑ 100%	All text fields normalized
Date Format Consistency	≥ 95%	☑ 97%	Validated with regex checks
Duplicate Record Removal	100%	☑ 100%	All duplicates removed using DISTINCT()
Referential Integrity (Joins across Tables)	≥ 85% match rate	☑ 88%	Learner–Opportunity–Cohort joins successful
Marketing Data Attribution	NA	✕ Not Available	No learner IDs in marketing dataset
Outlier Handling (Demographics)	Identified/Flagged	⚡ Flagged	Outliers flagged, preserved for authenticity

CHALLENGES VS RESOLUTIONS

CHALLENGES VS RESOLUTIONS

Challenge	Resolution
Missing data	NULLs preserved; critical fields flagg action
Duplicate entries	Cleaned using DISTINCT validation methods
Marketing disconnection	Separate dashboards with future integration
Text inconsistency	Casing normalized across datasets

IMPACT SUMMARY

IMPACT SUMMARY

BEFORE VS AFTER DATA TRANSFORMATION

BEFORE	AFTER
Fragmented and isolated datasets	Integraet Master Table covering 5 datasets
33%+ missing demographic data	Over 90% completeness achieved
Duplicate and inconsistent entries	Data standardized, dupiicates removed
Poor marketing attribution	Actionable insights for marketing and engagement
Manual and error-prone ETL processes	Automated, scalable ETL ready for real-time analytics

NEXT STEPS & FINAL TAKEAWAYS

Next Steps -

- Implement continuous Data Quality Monitoring pipelines.
- Strengthen marketing attribution models through improved tracking.
- Extend ETL automation to near real-time processing.
- Leverage cleaned datasets for Predictive Analytics and Machine Learning models.

Final Takeaways -

- The data environment is now trustworthy, consolidated, and strategically positioned for business intelligence.
- The ETL pipeline ensures scalability, flexibility, and sustainability.
- Key systemic data issues have been identified, addressed, and documented.
- Recommendations for marketing-data improvements are critical for full customer journey visibility.

THANK YOU