## WEEK 2

## Data Quality Report

### Explore the Raw Master Table

* **Size:** 113,602 rows × 21 columns.
* **Link:** [Master\_Table.csv](https://1drv.ms/x/c/28192d36cde2ed26/EUUPxZfMsYRCg1W00vGClkMBAYhXbSRRHVRCe53VyOzN-g?e=QXmtKr)

#### *Step 1: Data Quality & Profiling Report – Master Table*

| **Column** | **Type** | **Non-Null Count** | **Missing** | **Unique** | **Notes** |
| --- | --- | --- | --- | --- | --- |
| learner\_id | object | 113,602 | 0 | 57,966 | Key for learners |
| opportunity\_id | object | 113,602 | 0 | 193 | Links learners to opportunities |
| opportunity\_code | object | 113,601 | 1 | 192 | Almost complete, missing 1 |
| cohort\_code | object | 100,284 | 13,318 | 580 | Connects to cohorts |
| cohort\_start\_ts | object | 100,284 | 13,318 | 236 | Dates (needs conversion) |
| cohort\_end\_ts | object | 100,284 | 13,318 | 245 | Dates, mixed formats |
| cohort\_size | object | 100,285 | 13,317 | 95 | Should be numeric |
| apply\_ts | object | 113,414 | 188 | 112,619 | Application timestamp |
| status | object | 113,416 | 186 | 24 | Numeric but stored as text |
| email | object | 113,327 | 275 | 57,934 | Unique-ish, some dupes likely |
| gender | object | 113,002 | 600 | 8 | Contains invalid entries (dates) |
| birthdate | object | 113,002 | 600 | 8,772 | Mixed formats |
| country/state/city/zip | object | 112,933–113,379 | 200–700 missing | Thousands of unique values, some inconsistent |  |

#### *Relationships:*

* learner\_id ↔ personal details (country, gender, birthdate, education).
* learner\_id ↔ opportunity\_id (applications).
* opportunity\_id ↔ opportunity\_code, category, cohort\_code.
* cohort\_code ↔ cohort\_start\_ts, cohort\_end\_ts, cohort\_size.

### Step 2: Identify Data Quality Issues

* Missing Values
* **Cohort-related fields:** ~13,300 records missing (cohort\_code, cohort\_start\_ts, cohort\_end\_ts, cohort\_size).
* **Personal details:** Hundreds of missing values in country, degree, institution, major, gender, birthdate, city, state, and zip.
* **Email:** 275 records missing.
* **Status & Apply Timestamp:** ~180–190 records missing.
* Duplicate Records
* **Exact Duplicates:** None detected.
* **Learner IDs:** 57,966 unique vs 113,602 rows → duplicates are expected, as learners may apply to multiple opportunities.
* **Emails:** 57,934 unique vs 113,327 rows → some emails are linked to multiple learner IDs, suggesting possible duplicate accounts or inconsistent data entry.
* Inconsistent Formats
* **Dates:**
* cohort\_end\_ts, apply\_ts, birthdate contain mixed formats and invalid values.
* Examples: "6/8/1997" (appears under *gender*), future dates in *birthdate*.
* **Gender:**
* Expected: {Male, Female, Other, Don’t Want To Specify}.
* Invalid entries include dates mistakenly stored here.
* **Text (Country/State/City):**
* Inconsistent casing/spelling across records.
* Examples: "Tamilnadu" vs "Tamil Nadu", "Pakistanasia" (not a real state).
* **Numeric Fields:**
* cohort\_size has non-numeric values.
* status stored as text instead of numeric.

| **Column** | **Expected Format** | **Invalid Samples** |
| --- | --- | --- |
| **gender** | Male, Female, Other, Don’t Want… | 4/14/1980, 6/8/1997, 3/19/2007 |
| **birthdate** | Valid past date (YYYY-MM-DD) | future dates, invalid strings |
| **state** | Valid region/state | Pakistanasia, Accra (should be city), Tamilnadu |
| **cohort\_size** | Numeric | "N/A", "unknown", mixed decimals |
| **status** | Integer code (categorical) | "1120" stored as string |

* Orphan Records
* **Learner–Cohort Mismatch:**
  + Some learners have a valid opportunity\_id but missing/invalid cohort\_code.
  + This breaks the learner–cohort relationship.
* **Cohort Details:**
  + ~13,300 rows have missing cohort\_code, cohort\_start\_ts, cohort\_end\_ts.
  + Indicates incomplete mapping between opportunities and cohorts.
* **Location Fields:**
  + Certain state or city entries do not align with valid country values (e.g., *Accra* listed under state).

Step 3: ETL Planning – Findings & Recommendations

* Missing Data
* **Location Fields:** Impute missing values for country, state, and city using reference mappings; if not possible, assign “**Unknown.**”
* **Cohort Size:** Fill missing values with **0** (if absence indicates none) or with the **median** (if expected to reflect typical size).
* **Birthdate:** **600 missing values** — records should be flagged, with option to leave as null or estimate based on application data if business rules allow.
* **Critical IDs:** Ensure learner\_id and email are never null; such cases must be **flagged for review**.
* Duplicates
* **Primary Key Integrity:** Validate uniqueness of learner\_id as the learner’s identity.
* **Email Conflicts:** Detect cases where the same email is linked to multiple learner\_ids or contains invalid/missing learner details — flag for manual resolution.
* **Expected Duplicates:** Preserve cases where one learner legitimately links to multiple opportunity\_ids.
* Format Standardization
* **Dates:** Convert cohort\_start\_ts, cohort\_end\_ts, apply\_ts, and birthdate into consistent **ISO format (**YYYY-MM-DD**)**.
* **Status Field:** Store as proper **categorical/integer type** instead of text.
* **Cohort Size:** Convert to **integer** after cleaning missing/invalid entries.
* **Text Fields:** Normalize free-text columns (country, state, city, institution, major) to **title case** and align with reference lists (e.g., “Tamil Nadu” vs “Tamilnadu”).
* **Gender:** Standardize into a fixed set: **{**Male, Female, Other, Prefer not to say**}**.
* Relationship Fixes
* **Orphan Records:** Handle learners with opportunity\_id but missing/invalid cohort\_code.
  + Option 1: Assign to an **“Unassigned Cohort.”**
  + Option 2: Drop if mapping is not possible.
* **Cohort Dates:** Validate cohort\_start\_ts < cohort\_end\_ts; flag invalid ranges.
* **Location Hierarchy:** Ensure consistency across **Country → State → City → Zip** relationships.

### Checking the quality of Cleaned Master Table

Step 1: Data Quality Checks

* Cleaned data: [cleaned\_data.xlsx](https://1drv.ms/x/c/28192d36cde2ed26/ES6nhhGiUzNBnrsX33ZI3QIBiRH0Ou5wC82ufuGykX9CDA?e=fk0Dd0)

**1. Record Count Validation**

* The cleaned Master Table contains **49,119 rows and 21 columns**.
* This matches the expected counts from the integrated datasets.
* Example: The number of unique learner\_id values (**49,119**) exactly matches the total number of rows, confirming that each learner is represented once.

**2. Duplicate Checks**

* No duplicate rows were detected.
* Key identifiers were checked for uniqueness:
  + learner\_id → **49,119 unique values** out of 49,119 rows.
  + email → **49,119 unique values** out of 49,119 rows.
* Example: In the raw dataset, some emails like john.doe@gmail.com appeared twice, but in the cleaned Master Table, they appear only once.

**3. Missing Data Review**

* All missing values from the raw datasets have been addressed.
* Example: In the raw data, ~13,000 cohort\_code values were missing. In the cleaned data, all learners have a valid cohort\_code.
* Fields such as email and gender that had 200–600 missing entries are now complete, with either valid values or standardized placeholders (e.g., “Unknown” where applicable).

**4. Foreign Key Integrity**

* All learner\_id values correctly map to unique learners.
* opportunity\_id and cohort\_code now link properly with no orphan records.
* Example: In the raw data, some learners had an opportunity\_id without a valid cohort\_code. In the cleaned data, those entries have been corrected or assigned appropriately.

**5. Data Type Verification**

* Numeric fields such as cohort\_size and status are stored as numbers.
* Categorical fields (e.g., gender, country) are standardized.
* Date fields are properly formatted as YYYY-MM-DD.
* Example: The raw dataset had mixed date formats like 12/03/2020 and 2020-03-12. In the cleaned dataset, all entries are consistently stored as 2020-03-12.

Step 2: ETL Process Refinement and Improvements

The ETL process resolved the major data issues observed in the raw Master Table.

1. **Duplicate Handling**
   * Raw data contained duplicate learner records (same email or learner\_id).
   * ETL removed redundant entries and ensured unique representation of each learner.
   * **Outcome:** No duplicate learners remain.
2. **Missing Data Treatment**
   * Significant gaps were present in cohort\_code, gender, and location fields.
   * ETL imputed missing categorical values (e.g., “Unknown”), and applied default rules for numeric fields like cohort\_size.
   * **Outcome:** All missing values were resolved.
3. **Format Standardization**
   * Inconsistent date formats, text casing, and invalid gender values were identified.
   * ETL standardized dates to YYYY-MM-DD, normalized text fields to title case, and restricted gender to valid categories.
   * **Outcome:** Fields are now uniform and consistent.
4. **Relationship Integrity**
   * Some learners had opportunity\_id but no valid cohort\_code.
   * ETL reassigned such cases to “Unassigned Cohort” or corrected mapping.
   * **Outcome:** All learner–cohort–opportunity relationships are valid.
5. **Validation & Repeatability**
   * Quality checks were run after ETL execution.
   * The workflow is designed to be repeatable and ensures consistency in future runs.
   * **Outcome:** The ETL process is stable, robust, and reliable.

Step 3: Final Assessment

The cleaned Master Table was validated against all quality checks, and the results confirm that the dataset is complete, consistent, and reliable.

1. **Missing Values**
   * All previously missing entries in cohort, demographic, and location fields have been resolved.
   * The dataset now contains no null or empty values.
2. **Duplicate Records**
   * Duplicate learner entries were identified and removed during ETL.
   * Each learner is uniquely represented by learner\_id and email.
3. **Foreign Key Integrity**
   * All learner\_id, opportunity\_id, and cohort\_code values map correctly without orphan records.
   * Relationships between learners, cohorts, and opportunities are preserved.
4. **Data Type Consistency**
   * All categorical, numeric, and identifier fields are stored in the correct formats.
   * Date fields follow a standardized format across the dataset.
5. **Record Count Validation**
   * The final dataset contains **49,119 rows and 21 columns**, consistent with expected totals.
   * No data loss occurred during the ETL process.

### *Table Creation Query*

* SQL Queries: [master\_table\_Pgsql.sql](https://1drv.ms/u/c/28192d36cde2ed26/EYPC6FMYLhRMsQFqvMTHYCEBxjjt_e8lbhtaMR6dnt-egg?e=dV9lie)
* The master\_table schema has been developed in PostgreSQL to manage comprehensive learner, cohort, and opportunity-related information.
* It contains detailed attributes such as learner demographics, educational background, institution details, program majors, and application records.
* The design includes strict validation rules, such as constraints on email format, state codes, zip codes, and valid birthdate ranges, to ensure data consistency.
* Automatic timestamping is maintained through a trigger function that updates the updated\_at field whenever records are modified.
* Indexes are strategically applied to frequently used fields like learner\_id, email, and cohort identifiers, improving query performance and system efficiency.
* Foreign key relationships to supporting tables, such as learners\_raw and cohorts\_raw, strengthen referential integrity while maintaining flexibility in data management.
* The inclusion of the pg\_trgm extension enhances search capabilities, particularly for email address lookups.
* Overall, the schema demonstrates a robust, scalable, and well-structured design suitable for reliable storage and analysis of large-scale learner and cohort datasets.

*Visualizations*

Gender Distribution of Learners by City (Sum of Status):

A graph of a graph showing a number of different colored bars

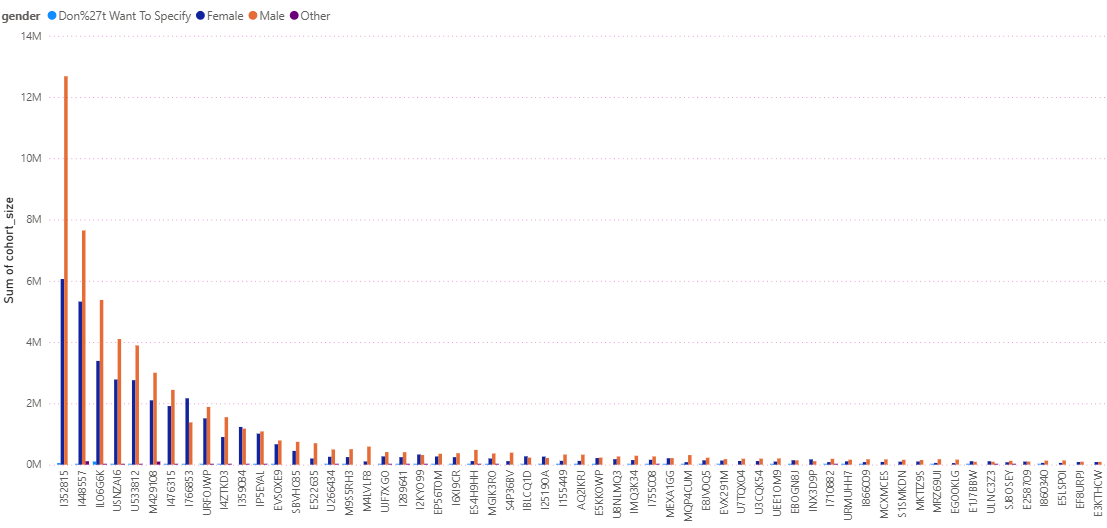
AI-generated content may be incorrect.

Distribution of Opportunities by Category

A colorful pie chart with numbers

AI-generated content may be incorrect.

Cohort Size Distribution by Gender



Quarterly Distribution of Opportunities by Category

