1 Module 2: Data Quality Adjustment

1.1 Background

Ensuring high-quality health service data is critical for accurate decision-making and analysis. However, routine health management information system (HMIS) data often contain *outliers* (extreme values due to reporting errors, data entry mistakes, or system issues) and missing values (due to incomplete reporting or data collection gaps). These issues can distort trends, mask true patterns, and significantly impact downstream analyses. The Data Quality Adjustment module addresses these challenges by applying systematic, evidence-based corrections to raw data.

1.2 Overview

This module adjusts facility-level service volumes using two methods:

1. Outlier Adjustment – replaces values flagged as outliers (`outlier_flag == 1`) using valid rolling or fallback averages. Only applied to non-exempt indicators.
2. Completeness Adjustment – replaces values from months flagged as incomplete (`completeness_flag != 1`) using recent valid (non-outlier, positive) values. Missing values (`NA`) are also adjusted.

Adjustments are applied across four scenarios:

Scenario	Outlier Adjustment	Completeness Adjustment
`none`	No	No
`outliers`	Yes	No
`completeness`	No	Yes
`both`	Yes	Yes

Each scenario produces a separate column: `count_final_none`, `count_final_outliers`, `count_final_completeness`, `count_final_both`

Adjustments are done at the facility × indicator × month level.

Some indicators including *Under 5 deaths* and *Maternal deaths* are excluded from all adjustments.

1.2.1 Key Features

- Dynamic adjustment logic that adapts to data availability
- Multiple fallback methods to ensure robust replacements
- Preservation of certain sensitive indicators (deaths) from adjustment
- Generation of facility, subnational, and national-level outputs
- Comprehensive tracking of adjustment methods used

1.3 Detailed Analysis Steps

1.3.1 Step 1: Data Preparation and Setup

- **Input files**: Raw HMIS data, outlier flags (from Module 1), and completeness data (from Module 1)
- **Exclusions**: Certain indicators (`u5_deaths`, `maternal_deaths`) are excluded from adjustment to preserve their original values
- **Working column**: The `count_working` column is initialized with actual service volumes (`count`) and serves as the target for all adjustments

1.3.2 Adjustment Logic and Scenarios

The adjustment logic is implemented through two functions:

1. `apply_adjustments()` defines the **adjustment rules** for replacing outliers and incomplete values.
2. `apply_adjustments_scenarios()` runs that logic across four scenarios: none, outliers only, completeness only, and both.

1.3.3 Step 2: Define Rules for Outlier Adjustment

Outlier adjustment is applied to any facility-month value flagged in Module 1 (`outlier_flag == 1`). The goal is to replace these outlier values using valid historical data from the same facility and indicator.

A rolling average is used to estimate expected values. A rolling average is the mean of a set of months around the target month. Only valid values are used—meaning the count must be greater than zero, not missing, and not flagged as an outlier.

The adjustment process follows this order:

1. Centered 6-Month Average (`roll6`)
   - Uses the three months before and three months after the outlier month
   - Provides a balanced average based on nearby trends
   - Applied when enough valid values exist on both sides of the month
   - Method tag**: `roll6`
2. Forward-Looking 6-Month Average (`fwd6`)
   - Used if the centered average can't be calculated (e.g. early in the time series)
   - Takes the average of the next six valid months
   - Method tag: `forward`
3. Backward-Looking 6-Month Average (`bwd6`)
   - Used if neither `roll6` nor `fwd6` are available
   - Takes the average of the six most recent valid months before the outlier
   - Method tag: `backward`
4. Same Month from Previous Year
   - If no valid 6-month average exists, the value from the **same calendar month in the previous year** is used (e.g., Jan 2023 for Jan 2024)
   - Only applied if that previous value is valid (not an outlier, and > 0)
   - Method tag: `same_month_last_year`
5. Median of All Historical Values
   - If all previous methods fail, the median of all valid historical values for that facility-indicator is used
   - Method tag: `fallback`

If no valid replacement can be found from any of these methods, the original outlier value is kept.

1.3.4 Step 3: Define Rules for Completeness Adjustment

Completeness adjustment is applied to any facility-month where the month is flagged as incomplete (`completeness_flag != 1`) in Module 1. The same rolling average logic is used, based only on valid historical values from the same facility and indicator.

The replacement follows this order:

1. Centered 6-Month Average (`roll6`)
   - Uses three valid months before and after the missing or incomplete month
   - Method tag: `roll6`
2. Forward-Looking 6-Month Average (`fwd6`)
   - Used if the centered average cannot be calculated
   - Method tag: `forward`

3. Backward-Looking 6-Month Average (`bwd6`)
    - Used if no centered or forward-looking values are available
    - Method tag: `backward`
4. Mean of All Historical Values
    - If no rolling averages can be calculated, uses the mean of all valid values for that facility-indicator
    - Method tag: `fallback`

If no valid replacement is found, the value remains missing.

1.3.5 Step 4: Scenario Processing

The module processes all four adjustment scenarios simultaneously:

1. None scenario: Original data with no adjustments
2. Outliers scenario: Outlier adjustment only
3. Completeness scenario: Completeness adjustment only
4. Both scenario: Sequential application of outlier then completeness adjustments

Each scenario produces a separate `count_final_[scenario]` column in the output.

1.3.6 Step 5: Geographic Aggregation

The module generates three levels of geographic aggregation:

1.3.6.1 Facility Level (`M2_adjusted_data.csv`)

- Individual facility data with all four adjustment scenarios
- Excludes `admin_area_1` to ensure schema consistency across levels

1.3.6.2 Subnational Level (`M2_adjusted_data_admin_area.csv`)

- Aggregated to subnational administrative areas (excluding national level)
- Sums facility-level adjusted values by geographic area, indicator, and time period
- Maintains all four adjustment scenarios

1.3.6.3 National Level (`M2_adjusted_data_national.csv`)

- Aggregated to national level (admin_area_1 only)
- Provides country-level totals for all indicators and time periods
- Includes all four adjustment scenarios for comparative analysis

---

Last edit 2025 September 3