Data Folder Dictionary

Understanding the Data Structure and Purpose

Document Purpose: This document explains the contents of the `data/` folder, describing what each variable means and the overall purpose of the data for future processing and analysis.

Folder Location: `/data/`

Folder Overview

The `data/` folder contains the raw and processed data files for the Tennessee Foster Care Youth Analysis project. This folder serves as the central repository for all data inputs and outputs used in the analysis.

File Contents

1. IL YOUTH DATASET FROM FY23-25.xlsx

File Type: Excel spreadsheet
Size: 33MB
Purpose: Primary source data containing raw assessment information for Tennessee foster care youth

What This File Contains:

- **Raw assessment data** from the Tennessee foster care system
- **Individual assessment responses** for each youth
- **Multiple assessment types** (CANS and LIFESKILL protocols)
- **Geographic and demographic information** for each youth
- **Assessment scores and indicators** across various domains

Data Structure:

- **Multiple sheets** containing different types of assessment data
- **Long format data** where each row represents an individual assessment response
- **Assessment indicators** covering various vulnerability domains
- **Scoring systems** for different types of assessments

---

2. bdaic created tables/

This subfolder contains processed and cleaned data tables created from the raw Excel file for analysis purposes.

2.1 df_youth.csv

File Type: CSV (Comma Separated Values)
Size: 411KB

Purpose: Youth-level summary table with one row per individual youth

What This File Contains:

- **Demographic information** for each youth
- **Geographic location data** (counties and zip codes)
- **Assessment summary scores** across vulnerability domains
- **Composite vulnerability measures** for analysis

Key Variables:

- **PERSON ID:** Unique identifier for each youth in the system
- **CURRENT AGE:** Age of the youth at time of assessment
- **GENDER:** Biological sex of the youth (MALE/FEMALE)
- **COMMITMENT COUNTY:** County where the youth is committed to foster care
- **RESPONSIBLE COUNTY:** County responsible for the youth's case management
- **REMOVAL ZIP CODE:** Zip code where the youth was removed from their home
- **PLACEMENT ZIP CODE:** Zip code where the youth is currently placed
- **LOCATION BEGIN DATE:** Start date of the current placement
- **LOCATION END DATE:** End date of the current placement (if applicable)
- **CANS:** Summary score from CANS assessment protocol
- **LIFESKILL:** Summary score from LIFESKILL assessment protocol

Data Characteristics:

- **One row per youth** (denormalized structure)
- **Geographic identifiers** for mapping and regional analysis
- **Temporal data** for tracking placement changes over time
- **Assessment summaries** for vulnerability analysis

2.2 df_assessment_questions.csv

File Type: CSV (Comma Separated Values)
Size: 120MB
Purpose: Detailed assessment-level data with individual question responses

What This File Contains:

- **Individual assessment responses** for each youth
- **Question-level detail** from assessment protocols
- **Scoring information** for specific assessment indicators
- **Assessment metadata** (dates, types, administrators)

Key Variables:

- **PERSON ID:** Links to the youth in df_youth.csv
- **ASSESSMENT TYPE:** Type of assessment (CANS, LIFESKILL, etc.)
- **ASSESSMENT INDICATOR:** Specific question or indicator being assessed
- **SCORE-RESULT:** Numerical score for the assessment indicator
- **GENERAL RESULT DESCRIPTION:** Text description of the assessment result

- **Assessment dates and administrative information**

Data Characteristics:

- **Multiple rows per youth** (one per assessment question)
- **Detailed scoring** for vulnerability domain analysis
- **Assessment protocol information** for methodology validation
- **Raw response data** for custom scoring algorithms

Data Relationships and Structure

Data Flow

1. **Raw Data:** `IL YOUTH DATASET FROM FY23-25.xlsx`
2. **Processing:** Data cleaning and transformation
3. **Output Tables:** `df_youth.csv` and `df_assessment_questions.csv`

Table Relationships

- **df_youth.csv** contains **one row per youth** with summary information
- **df_assessment_questions.csv** contains **multiple rows per youth** with detailed assessment data
- **PERSON ID** serves as the primary key linking the tables
- **Geographic fields** enable spatial analysis and mapping

Data Granularity

- **Youth Level:** Demographics, location, summary scores
- **Assessment Level:** Individual question responses and detailed scoring
- **Geographic Level:** County and zip code identifiers for spatial analysis

---

Variable Definitions and Purpose

Demographic Variables

- **PERSON ID:** Unique identifier for data linking and deduplication
- **CURRENT AGE:** Age-based vulnerability analysis and service planning
- **GENDER:** Gender-specific vulnerability patterns and service needs

Geographic Variables

- **COMMITMENT COUNTY:** Primary jurisdiction for foster care services
- **RESPONSIBLE COUNTY:** Administrative responsibility for case management
- **REMOVAL ZIP CODE:** Geographic origin of the youth's situation
- **PLACEMENT ZIP CODE:** Current location for service delivery planning

Temporal Variables

- **LOCATION BEGIN DATE:** Start of current placement for duration analysis
- **LOCATION END DATE:** Placement stability and transition planning

Assessment Variables

- **CANS:** Child and Adolescent Needs and Strengths assessment summary
- **LIFESKILL:** Life skills assessment for transition planning
- **Individual assessment indicators:** Detailed vulnerability domain scoring

---

Data Purpose and Use Cases

Primary Analysis Goals

1. **Geographic Distribution Analysis:** Understanding where youth are located across Tennessee
2. **Vulnerability Assessment:** Identifying needs across housing, mental health, and relationship domains
3. **Service Planning:** Informing resource allocation and service delivery strategies
4. **Policy Development:** Evidence-based recommendations for foster care system improvement

Analytic Capabilities

- **Spatial Analysis:** County and zip code level vulnerability mapping
- **Demographic Analysis:** Age and gender-based vulnerability patterns
- **Temporal Analysis:** Placement stability and transition timing
- **Domain-Specific Analysis:** Housing, mental health, and relationship vulnerability assessment

Output Applications

- **Geographic Vulnerability Maps:** Visual representation of need across Tennessee
- **County-Level Summaries:** Regional vulnerability profiles for service planning
- **Youth-Level Profiles:** Individual vulnerability assessments for case management
- **Policy Recommendations:** Data-driven guidance for system improvement

---

Data Quality and Considerations

Data Strengths

- **Comprehensive Coverage:** 96 counties across Tennessee represented
- **Standardized Protocols:** CANS and LIFESKILL assessment methodologies
- **Geographic Detail:** County and zip code level location data
- **Temporal Consistency:** FY23-25 data period for trend analysis

Data Limitations

- **Missing Values:** Some geographic fields may be incomplete
- **Assessment Timing:** Varying frequencies of assessment administration
- **Geographic Accuracy:** Zip code and county boundary considerations
- **Data Completeness:** Some youth may have partial assessment data

Processing Considerations

- **Data Cleaning:** Handling missing values and data inconsistencies

- **Geographic Standardization:** County name and zip code validation
- **Assessment Scoring:** Standardizing different assessment protocols
- **Data Aggregation:** Creating summary measures from detailed responses

Future Processing Guidelines

Data Loading

- Use appropriate data types for each variable (string for IDs, numeric for scores)
- Handle missing values appropriately for each variable type
- Validate geographic identifiers against Tennessee county and zip code lists

Data Transformation

- Create vulnerability domain scores from individual assessment indicators
- Aggregate youth-level summaries from assessment-level data
- Generate geographic summaries at county and zip code levels

Quality Assurance

- Validate PERSON ID uniqueness across tables
- Check geographic field consistency and completeness
- Verify assessment score ranges and validity
- Ensure temporal data logical consistency

Analysis Preparation

- Create derived variables for vulnerability domain analysis
- Generate geographic identifiers for spatial analysis
- Prepare demographic breakdowns for subgroup analysis
- Structure data for statistical modeling and visualization

File Access and Usage

File Formats

- **Excel (.xlsx):** Raw data source, requires Excel or compatible software
- **CSV (.csv):** Processed data tables, accessible in most data analysis tools

Software Compatibility

- **Python:** pandas, openpyxl for data loading and processing
- **R:** readxl, readr for data import and manipulation
- **Excel:** Direct access to raw data and processed tables
- **GIS Software:** Geographic data for mapping and spatial analysis

Data Security

- **Confidentiality:** Youth identifiers should be handled according to privacy protocols

- **Access Control:** Limit access to authorized personnel and systems
- **Data Sharing:** Aggregate data appropriately for external reporting
- **Retention:** Follow data governance policies for long-term storage

---

Summary

The `data/` folder contains the foundational data for understanding Tennessee's foster care youth landscape. The raw Excel file provides comprehensive assessment data, while the processed CSV tables offer cleaned and structured data for analysis. These files enable geographic vulnerability analysis, demographic pattern identification, and evidence-based policy development for the foster care system.

Key Takeaway: This data structure supports comprehensive analysis across multiple dimensions (geographic, demographic, temporal, and vulnerability domains) while maintaining the flexibility to address specific research questions and policy needs.