Using Open-Source and SAS Together for Data Quality

Python Meets the SAS QKB

Mickey Schauf, Sr. Product Manager

sas innovate 2026

Why Data Cleansing Is A Problem

A Quick Look at SAS Methods for Data Cleansing

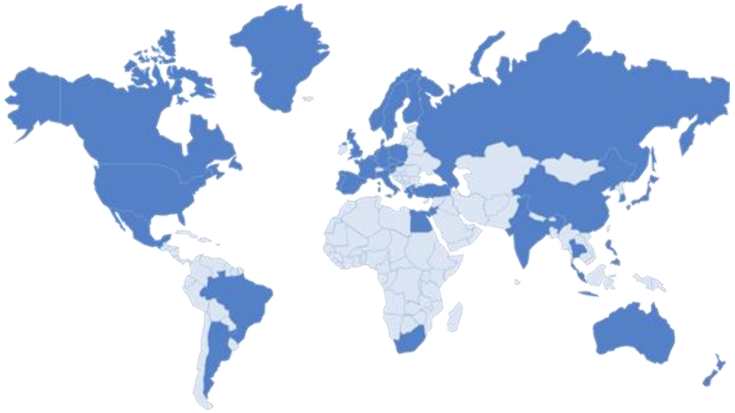
SAS Data Cleansing Methods

Category	Purpose	Typical SAS Tools	Example Techniques
1. Standardizing Values	Ensure consistent formats	DATA step, DQSTANDARDIZE	Case normalization, date standardization, remove punctuation
2. Parsing & Tokenizing	Break fields into components	DQPARSE, DATA step	Parse names, addresses, split strings
3. Data Type Corrections	Fix incorrect types	INPUT/PUT, Informat/Format	Char-to-numeric conversion, date corrections
4. Deduplication & Entity Resolution	Identify and merge duplicates	DQMATCH, COMPGED	Fuzzy matching, clustering duplicates
5. Validation Rules & Constraints	Enforce business rules	DATA step logic, DataFlux Rules	Range checks, conditional validation, domain rules
6. Data Enrichment	Add missing or better data	Lookup tables, DataFlux	Geocoding, reference-data enhancement
7. Name & Entity Cleansing	Normalize and validate personal info	DQSTANDARDIZE, DQPARSE	Name parsing, gender identification, entity normalization
8. Text Cleansing	Clean unstructured fields	DATA step, text routines	Remove punctuation, normalize strings, tokenization
9. Schema & Metadata Alignment	Fix structure inconsistencies	SAS DI Studio, Viya Pipelines	Metadata enforcement, field alignment, type validation

What is SAS’s Quality Knowledge Base

Why It’s Important

Common Data Quality Rules for...



47 countries

&

36 languages

over 51 locales, including different writing systems like Arabic, Kanji, Greek, Hebrew, Chinese ...

- **Parsing**
- **Standardization**
- **Matching**
- **Entity Resolution**
- **Identification**
- **Gender Analysis**
- **Enrichment**
- **Pattern Analysis**
- **Casing**
- **Extraction**
- **Locale Guess**
- **Language Guess**

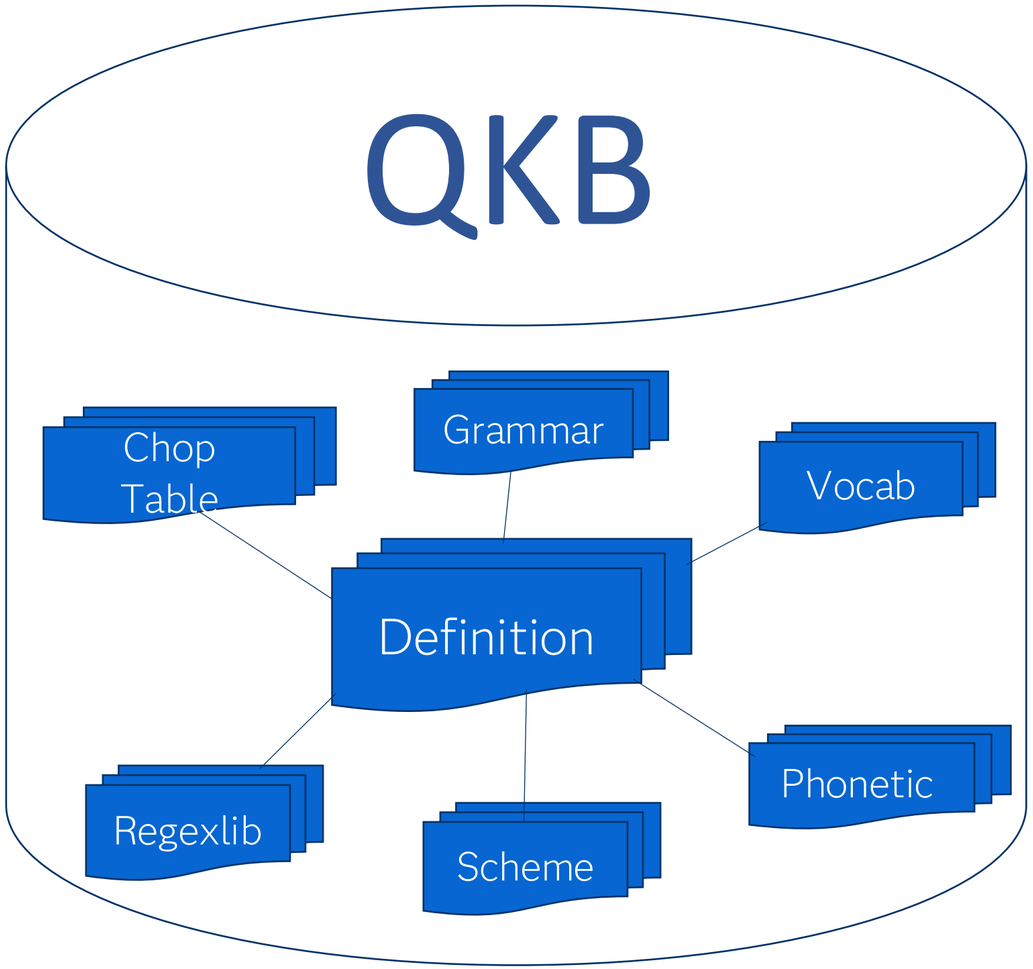
115

Data Types

3rd Party & Product Data Types

Related Data Types, such as Persons, Organizations, Vendors, Citizens, Suppliers, Patients, Vehicles, Product, Size, Colors ...

Architecture of the Quality Knowledge Base



A Definition is like a program.

- Consists of parameter values and pointers to specific library files that enable the definition to produce the desired output
- Definitions are used by DQ Functions
- Some definitions are used by other definitions

Chop Tables, Grammars, Vocabularies, Regexlibs, Schemes, and Phonetic Libraries

- Collectively called libraries
- Contain data and logic that are used by one or more definitions

The Quality Knowledge Base is AI

- Expert Systems were among the first successful forms of Artificial Intelligence (AI) software.
- An Expert System is a program that emulates the decision-making ability of a human expert.
- Facts and rules used by an Expert System are contained in a Knowledge Base.

Demo

SAS Calls to the QKB vs. Python Calls to the QKB

Why This Matters

What the QKB allows Customers to do:

- Impact of democratized business rule creation
  - The QKB allows users to build data quality processes using SAS or Python and it ensures predictable standardized output.
  - Language and location understanding will allow QKB users to become proficient much more quickly.
- Embracing the open-source community
  - Allowing programmers to use the environment & language they prefer helps with adoption which will ultimately help with teams across the organization becoming more standardized.
- Empowered workforce
- Stay competitive in a data-driven landscape

Thank You for Your Time Today!

SAS Quality Knowledge Base

sas innovate 2026

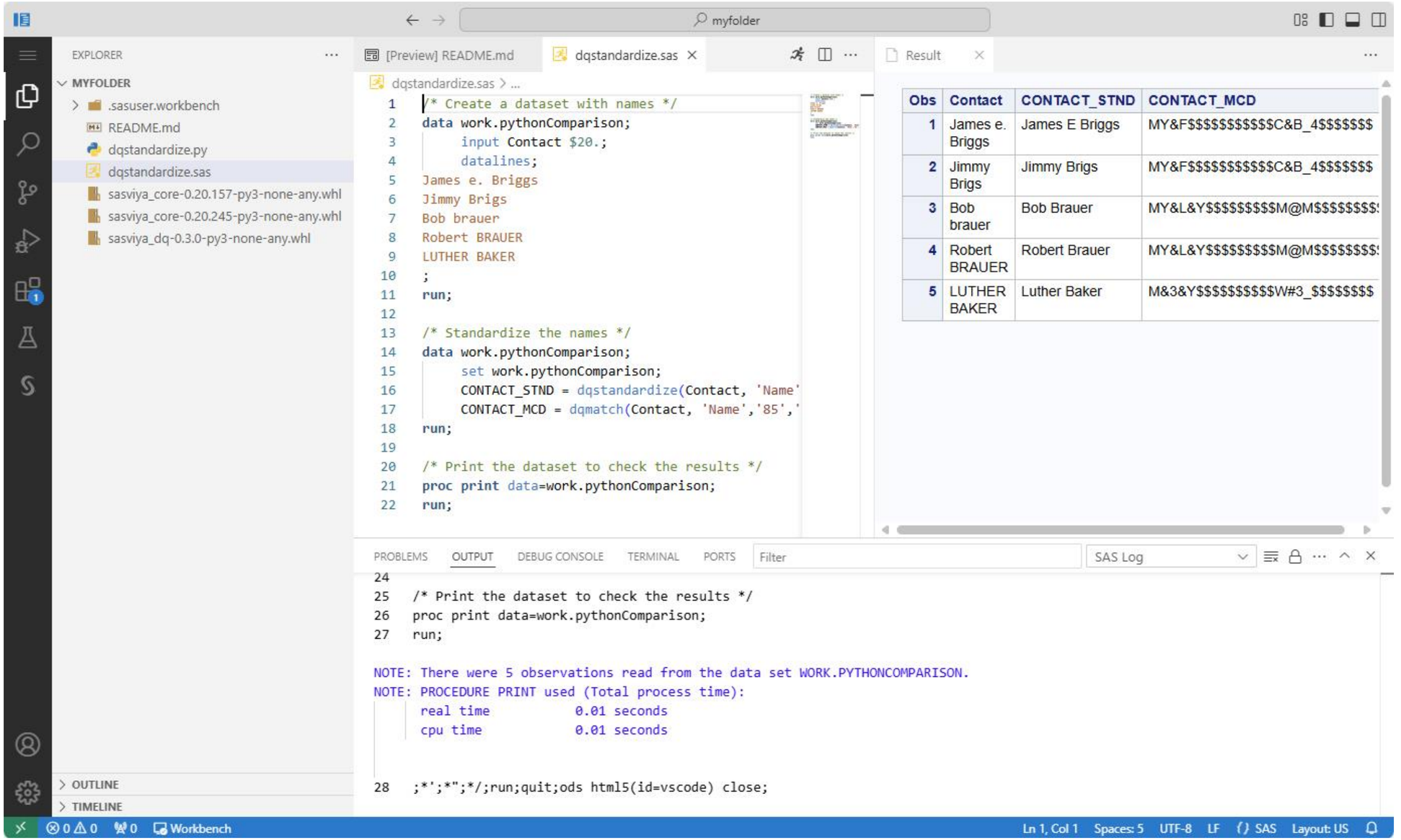
Addendum

Why It’s Important

SAS Data Cleansing Methods: Comparison Table

Category	Purpose	Typical SAS Tools	Example Techniques
1. Handling Missing Data	Improve completeness	Base SAS, PROC MI	Imputation, remove missing, analyze missing patterns
2. Standardizing Values	Ensure consistent formats	DATA step, DQSTANDARDIZE	Case normalization, date standardization, remove punctuation
3. Parsing & Tokenizing	Break fields into components	DQPARSE, DATA step	Parse names, addresses, split strings
4. Data Type Corrections	Fix incorrect types	INPUT/PUT, Informat/Format	Char-to-numeric conversion, date corrections
5. Deduplication & Entity Resolution	Identify and merge duplicates	DQMATCH, COMPGED, SORT NODUPKEY	Fuzzy matching, clustering, duplicate removal
6. Outlier & Anomaly Detection	Remove abnormalities	PROC UNIVARIATE, VDMML	Statistical outliers, ML anomaly detection
7. Validation Rules & Constraints	Enforce business rules	DATA step logic, DataFlux Rules	Range checks, conditional validation, domain rules
8. Data Profiling	Understand data quality issues	SAS Data Quality, PROC FREQ	Frequency patterns, completeness, uniqueness
9. Data Enrichment	Add missing or better data	Lookup tables, DataFlux	Geocoding, reference-data enhancement
10. Address Verification / Cleansing	Standardize & validate addresses	DataFlux Address Verification	Postal corrections, DPV/CASS, standard formatting
11. Name & Entity Cleansing	Normalize and validate personal info	DQSTANDARDIZE, DQPARSE	Name parsing, gender identification, entity normalization
12. Text Cleansing	Clean unstructured fields	DATA step, text routines	Remove punctuation, normalize strings, tokenization
13. Schema & Metadata Alignment	Fix structure inconsistencies	SAS DI Studio, Viya Pipelines	Metadata enforcement, field alignment, type validation
14. Transformation-Based Cleansing	Improve usability of data	PROC SQL, DATA step	Aggregation, merging, pivots, conditional transformations
15. Workflow-Based Cleansing	End-to-end cleansing	SAS Data Management Studio, Viya Pipelines	Profile → standardize → match → validate workflows

SAS Workbench using SAS



SAS Workbench using Python