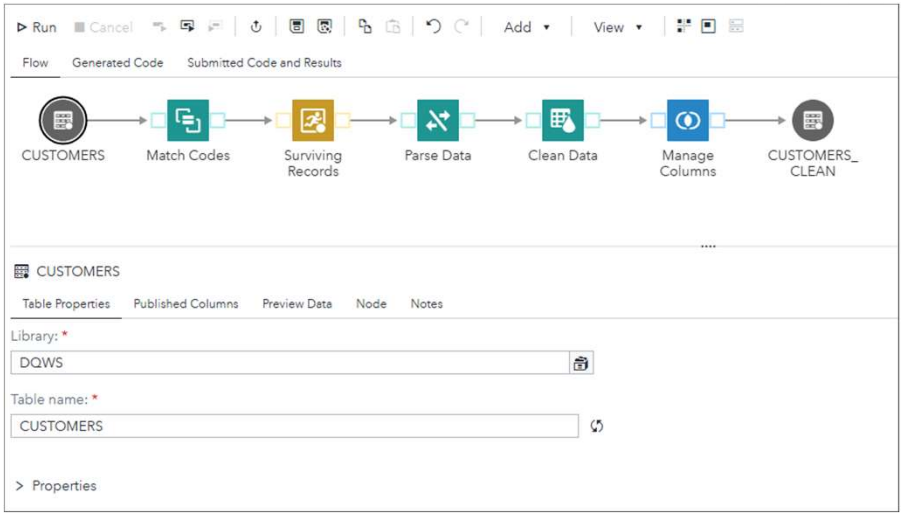
sas innovate 2025

Power Washing Your Data: Using SAS Data Quality Steps in SAS Studio Flows

Hands-on Workshop

Grace Barnhill, Technical Training Consultant
SAS EDU Data Management Content Development

SAS Studio Flows



A *flow* is a sequence of operations on data. Data and operations are represented in SAS Studio by steps. Each step in a flow is represented by a node on the flow canvas. The output of one node can be the input to another node.

As you build a flow, SAS Studio automatically generates SAS code for each node. You can use flows to prepare data for reporting and analysis.

Data Scenario: Customers

CUSTOMERS — Table rows: 60 | Columns: 7 of 9 | Rows 1 to 60

Enter expression

	CustomerID	Name	Gen...	Address1	Birthday	Email
1	.	AGNES R. YATES		4334 CHESTNUT ST,POLK CITY,FLORIDA,...	.	
2	.	JAMES ATTWOOD		3238 HERITAGE RD,SAN JOAQUIN,CALIF...	.	
3	100000014	Agnes R. Yates		4334 Chestnut Street,Polk City,FL,33868,US	.	AgnesRYates@pookmail.com
4	100000012	Anna R. Christian		3323 Morgan Street,Fort Walton beach,FL...	.	AnnaChristi@dodgit.com
5	100000010	Attwood, James	M	3238 Heritage Road,San Joaquin,CA,936...	1988-06-29	
6	.	CALVIN WENZEL		1007 THOMPSON DR,DUBLIN,CALIFORN...	.	
7	.	CATHERINE M. THOMAS		3005 BROOKSIDE DRIVE,GUIN,ALABAM...	.	
8	.	CHRISTIAN, ANNA		3323 MORGAN ST,FORT WALTON BEAC...	.	
9	.	CURRY, KRISTI R.		399 MASONIC HILL ROAD,LITTLE ROCK,...	.	
10	100000007	Calvin B. Wenzel		1007 Thompson drive,Dublin,CA,94568,US	.	CalWenzel@mailinator.com
11	100000018	Carolan, Kris J.	M	3935 Lakeland Park Driv,Duluth,GA,3009...	1986-09-23	
12	100000001	Catherine Thomas		3005 Brookside Drive,Guin,AL,35563,US	.	KatherineMThomas@mailinator.co

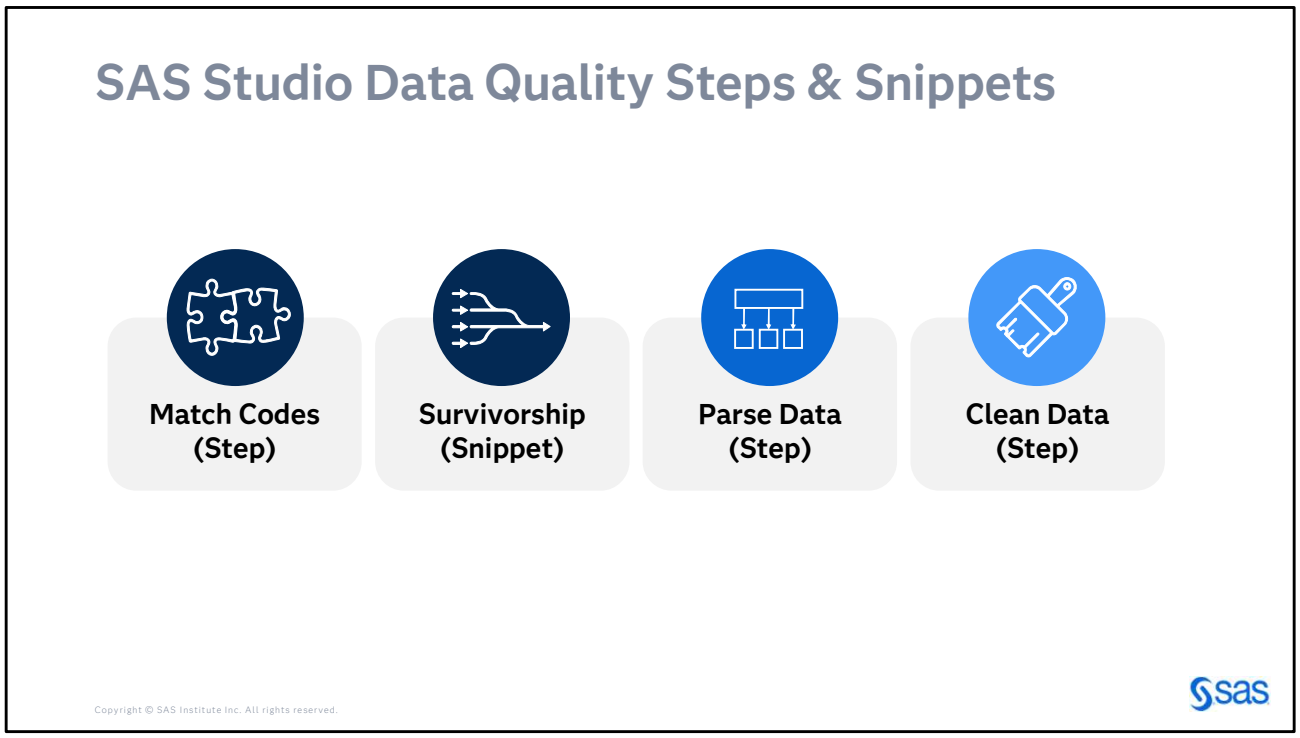
In this workshop, we'll be using the **CUSTOMERS** table, which has many data quality issues. Note that:

- Each customer appears three times.
- All rows have *Name* and *Address* values.
- Most rows have a non-missing *CustomerID* value.
- Customer *Name* values differ in each appearance (casing, name order, inclusion of middle initial or use of nickname).
- Rows with non-missing *Gender* values also have non-missing *CustomerID*, *Birthday*, *Occupation*, and *Company* values.
- Rows with non-missing *Email* values also have non-missing *CustomerID* and *Phone* values.

Data Scenario: Customers

1. Combine duplicate records (put all non-missing values in one record)
2. Parse dense variables into separate columns (i.e. **Address → Street, City, State, Zip, Country**)
3. Standardize final values

We'll follow three main steps to clean the **CUSTOMERS** table.
First, we'll combine the duplicate records. We'll put all non-missing values in one record to create a completed row.
Next, we'll parse out the variables that hold several pieces of information, like **Address**.
Lastly, we'll standardize the final values.



To complete our goals, we'll use the Match Codes step, the Survivorship code snippet, the Parse Data step, and the Clean Data step.

SAS Quality Knowledge Base (QKB) Overview

The SAS Quality Knowledge Base (QKB) is a collection of files that store rules, logic, and reference data used in data management and data quality operations.

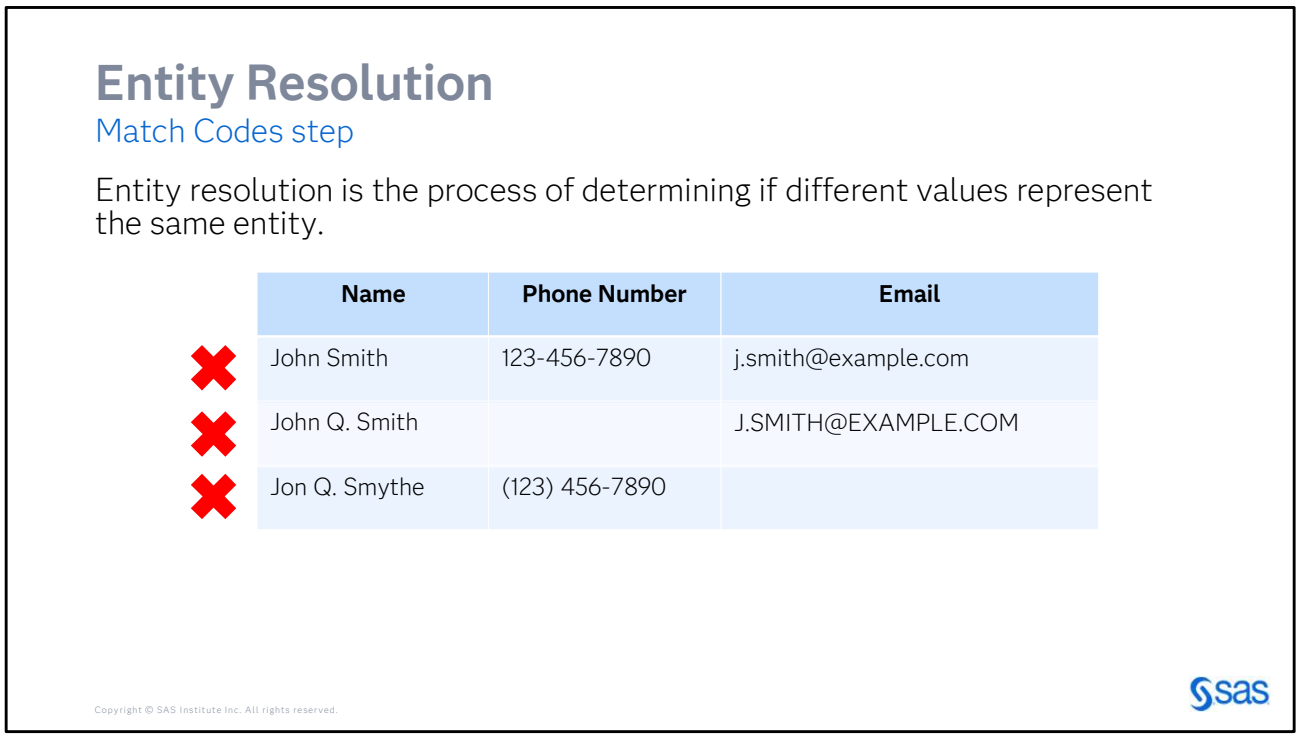
Input Value	QKB definition	Output value
ms jane smith	"Proper (Name)" Case	Ms Jane Smith
	"Name" Parse	Prefix: ms Given Name: jane Family Name: smith
	"Name" Gender Analysis	F
	"Field Content" Identification Analysis	INDIVIDUAL

Before we learn how to use these steps, let's discuss the magic underlying them: the SAS Quality Knowledge Base.

The SAS Quality Knowledge Base or (QKB) is a collection of files that store rules, logic, and reference data used in data management and data quality operations.
The QKB for Contact Information contains tools for performing standard data quality operations on contact data like individual and organization names, addresses, job titles, phone numbers, email addresses, and more.

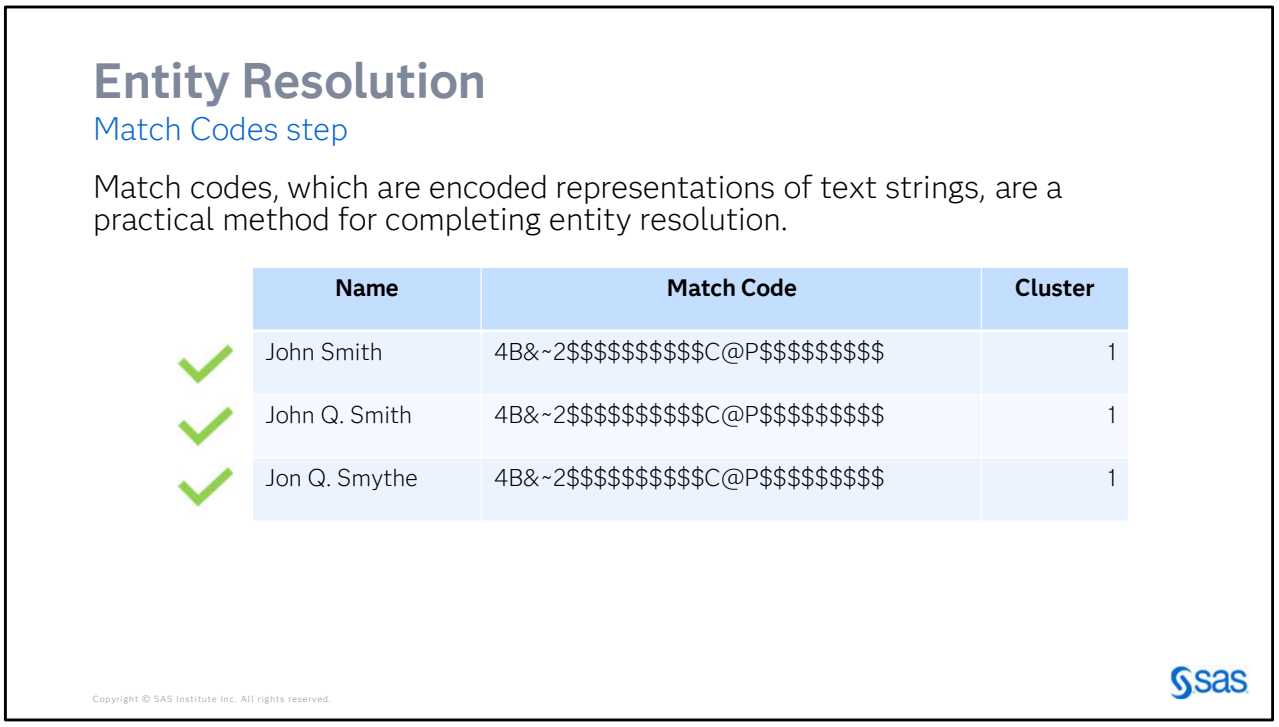
Forexample, let's see what the QKB can do with the value "ms jane smith", all lowercase. You can use the **Proper (Name)** case definition to improve the value's casing. You can also use the **Name** parse definition to identify the different tokens in this value: the prefix *Ms*, given name *Jane*, and family name *Smith*.
If you want to dig deeper, you can use analysis definitions to gather some metadata on this value. The **Name** gender analysis definition tells you that this name is likely female, while the **Field Content** identification analysis definition tells you that this value represents an individual person.

This is a small example of what you can achieve with the QKB.



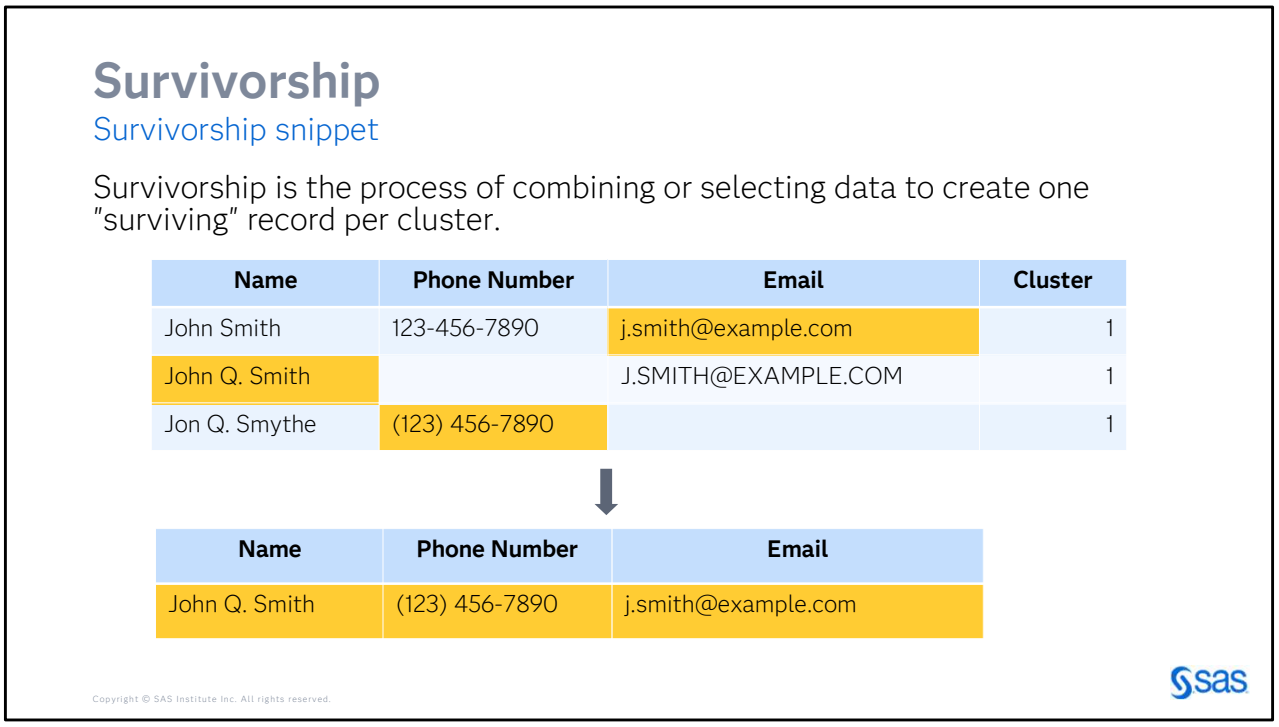
Entity resolution is the process of determining if different values represent the same entity. Specifically, this can mean finding multiple occurrences of the same entity within one or more data sets. Match codes are a practical method for completing entity resolution.

Say that you're looking for all occurrences of the name John Smith in your data. There might be slight differences in each occurrence of the name. For example, a middle initial might be included, or the name might have an alternate spelling.



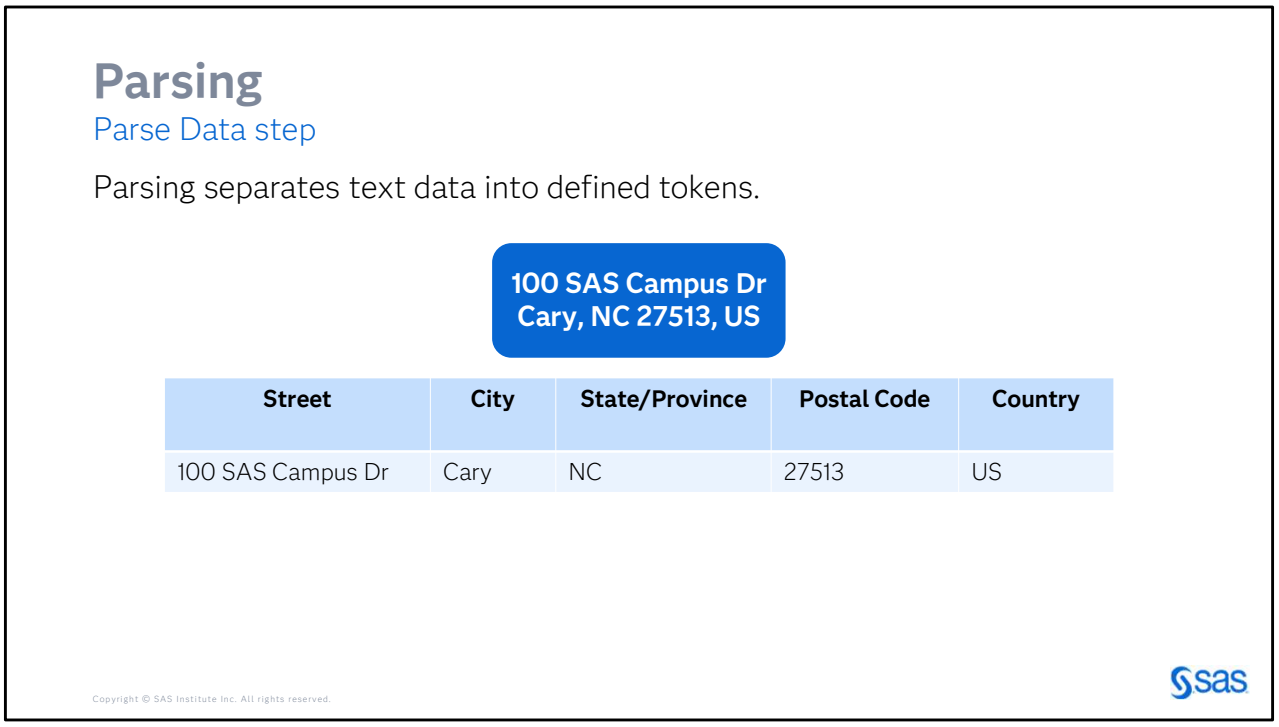
Match codes are an encoded representation of a text string. Match codes are generated based on the data type, the tokens in the string, and the sensitivity level. Generating match codes for the three different values of John Smith yields the same result.
This means the computer can now recognize that these values represent the same person, even though they are not exactly the same.

In addition, data can be clustered or grouped based on equivalent match codes.



Survivorship is the process of combining or selecting data to create one "surviving" record per cluster.

Rules are set to select which values to retain for specific variables. The output is a single completed record.



Parsing separates text data into defined tokens.

For example, parsing an address would return the **Street**, **City**, **State/Province**, **Postal Code**, and **Country** tokens.

Standardization

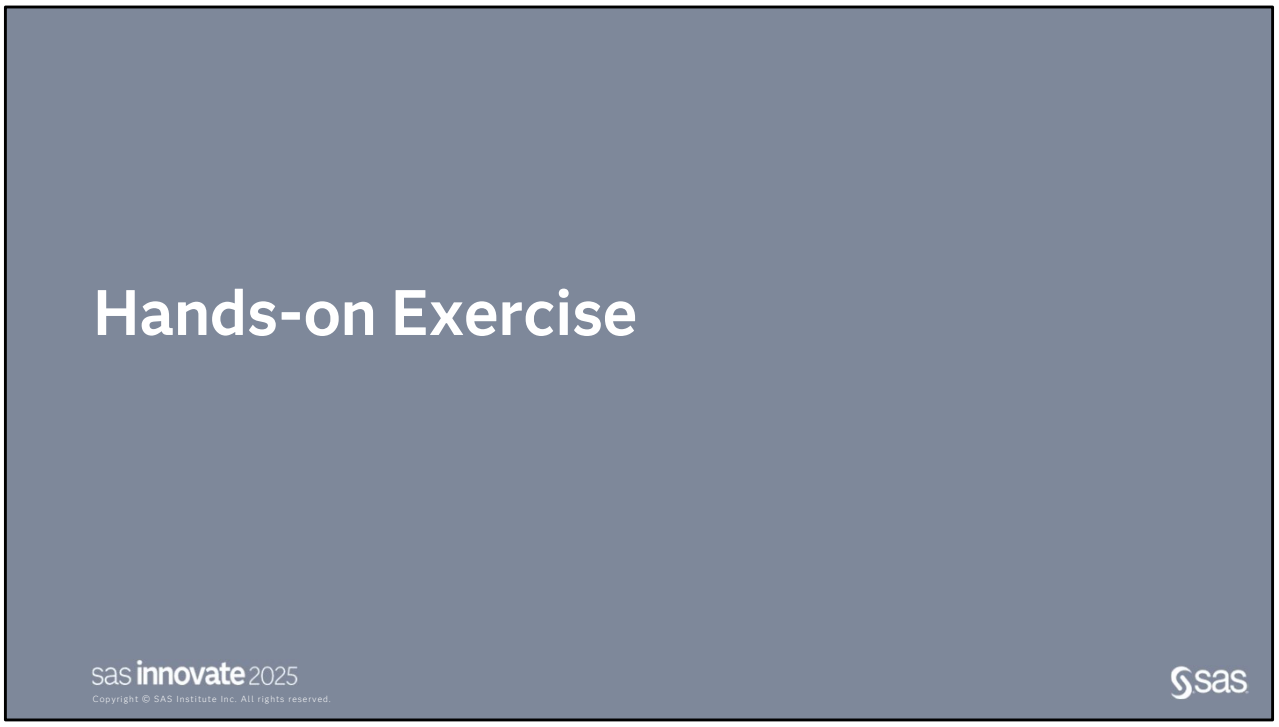
Clean Data step

Standardization definitions tokenize input strings and transform each token separately to meet a defined standard.

Original Value	Standardization definition (ENUSA)	Standardized Value
cary nc 27513	City – State/Province – Postal Code	Cary, NC 27513
04 July 2005	Date (MDY)	07/04/2005
ACME management grp	Organization	Acme Mgmt Group
123 456 78 9 0	Space Removal	1234567890

Standardization definitions tokenize input strings and transform each token separately to meet a defined standard.

Standardization definitions are available for dozens of semantic data types and data quality tasks. For example, you can standardize the second line of an address with the **City – State/Province – Postal Code** definition. You can also standardize a set of dates with the **Date (MDY)** definition. The **Organization** definition is helpful for standardizing company and organization names, especially when you aren't sure which parts of the name can be shortened. You can even do some simple data cleanup with definitions like **Space Removal**, which simply removes all spaces from the input string.



Hands-on exercise is available at https://github.com/SAS-Innovate-2025/Power-Washing-Your-Data-Using-SAS-Data-Quality-Steps-in-SAS-Studio-Flows/.

Thank you for attending this workshop!

sas innovate 2025