

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

---

- [Exercise Description](#)
- [Log in to SAS Viya](#)
- [Create a SAS Studio Flow](#)
  - [Generate Match Codes](#)
  - [Remove Duplicates with Survivorship Snippet](#)
  - [Parse Address Data](#)
  - [Standardize Data](#)
  - [Create Output Table](#)
- [Exercise Completed](#)

## Exercise Description

Update your data cleaning process with point-and-click data quality steps in SAS Studio Flows! In this hands-on workshop, you'll learn how to build a flow that deduplicates, parses, and standardizes your data with SAS Data Quality steps and code snippets.



## Log in to SAS Viya

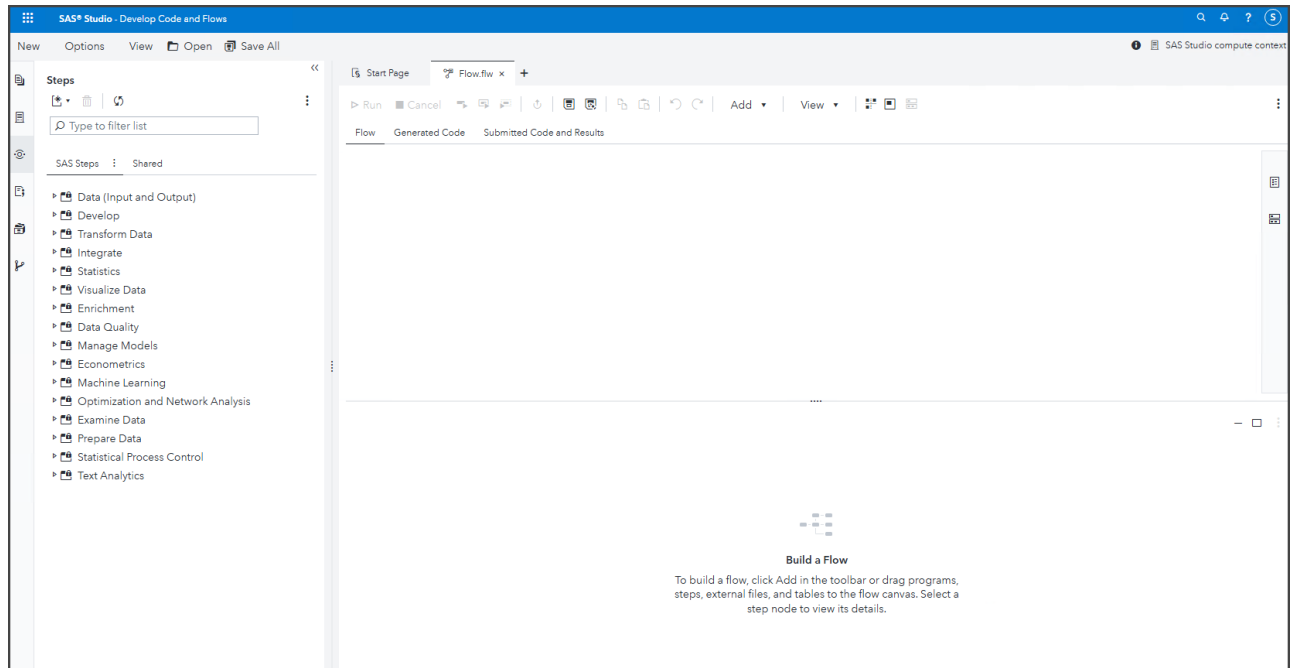
Open a new window in the *Google Chrome* browser and select the **SAS Viya** bookmark.

- ID: **student**
- Password: **Metadata0**

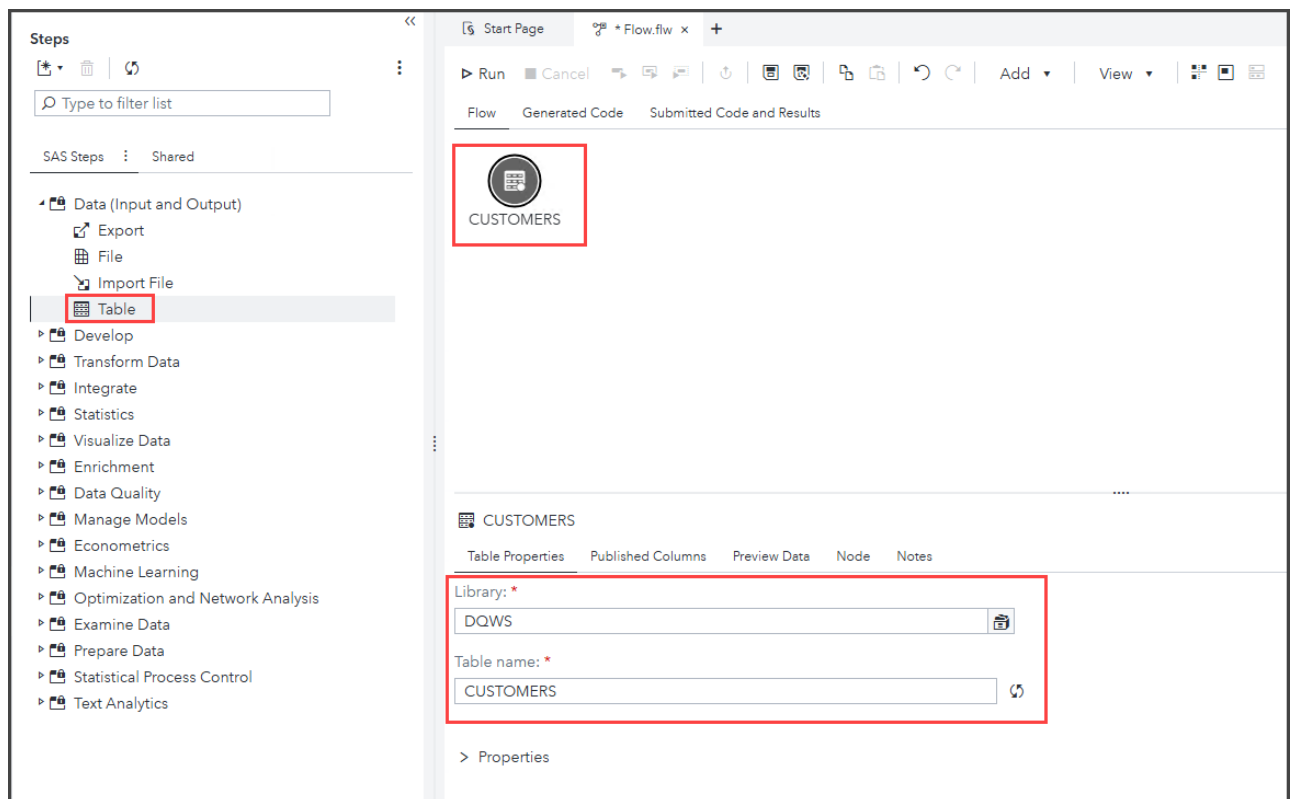
Select **No** when prompted about accepting *Admin* privileges.

## Create a SAS Studio Flow



1. Select  → **Develop Code and Flows** to open *SAS Studio*.
2. Select **New** → **Flow**.
3. Select  to view the **Steps** pane.



4. Expand the *Data (Input and Output)* section of the *Steps* pane. Double-click the **Table** step to add it to the flow canvas.
5. Select the **Table** node on the flow canvas. In the **Table Properties** section, select the following:
  - Library: **DQWS**
  - Table name: **CUSTOMERS**



6. Click **Preview Data** to view a subset of rows from the table.

If needed, click  to hide the Steps pane, or click  to maximize the preview.

Enter expression									
	CustomerID	Name	Gender	Address1	Birthday	Occupation	Company	Email	Phone
1	.	AGNES R. YATES		4334 CHESTNUT ST,POLK,CITY,FLORID...	.				
2	.	JAMES ATTWOOD		3238 HERITAGE RD,SAN JOAQUIN,CAL...	.				
3	100000014	Agnes R. Yates		4334 Chestnut Street,Polk City,FL,33868...	.			AgnesRYates@poo...	727-460-5745
4	100000012	Anna R. Christian		3323 Morgan Street,Fort Walton beach,...	.			AnnaChrist@dodg...	850-467-1006
5	100000010	Attwood, James	M	3238 Heritage Road,San Joaquin,CA,93...	1988-06-29	Information sy...	Thom McAn S...		
6	.	CALVIN WENZEL		1007 THOMPSON DR,DUBLIN,CALIFO...	.				
7	.	CATHERINE M. THOMAS		3005 BROOKSIDE DRIVE,GUIN,ALABA...	.				
8	.	CHRISTIAN, ANNA		3323 MORGAN ST,FORT WALTON BEA...	.				
9	.	CURRY, KRISTI R.		399 MASONIC HILL ROAD,LITTLE ROC...	.				
10	100000007	Calvin B. Wenzel		1007 Thompson drive,Dublin,CA,94568...	.			CalWenzel@mailin...	510-453-8297
11	100000018	Carolann, Kris J.	M	3935 Lakeland Park Driv,Duluth,GA,300...	1986-09-23	Real property ...	Dun Rite Lawn...		
12	100000001	Catherine Thomas		3005 Brookside Drive,Guin,AL,35563,US	.			KatherineMThoma...	205-412-9602
13	100000013	Cheeks, Linda	F	3385 Rinehart Road,Miami,FL,33128,US	1999-09-30	Delivery servi...	Home Centers		
14	100000012	Christian, Anna R.	F	3323 Morgan Street,Fort Walton beach,...	1979-07-30	Mixing and bl...	Sambo's		
15	100000004	Curry, Kristi	F	399 Masonic Hill Road,Little Rock,AR,72...	1996-05-13	Industrial pro...	Giant		
16	.	DICK REYES		4819 KERRY WAY,LOS ANGELES,CALIF...	.				


CUSTOMERS has **60** rows and **9** columns. Note that:

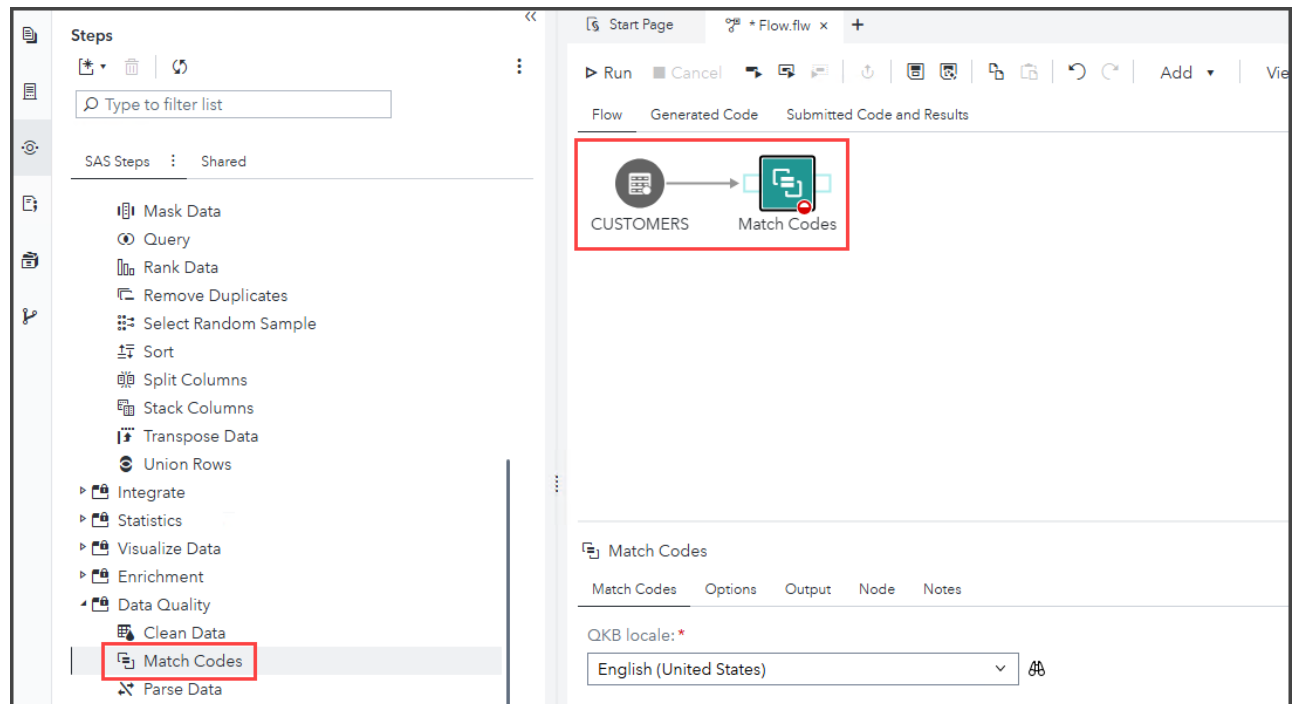
- Each customer appears three times.
- All rows have *Name* and *Address* values.
- Most rows have a non-missing *CustomerID* value.
- Customer *Name* values differ in each appearance (casing, name order, inclusion of middle initial or use of nickname).
- Rows with non-missing *Gender* values also have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

We would like to deduplicate this data by combining unique values into one row per customer. First, we'll generate match codes to determine which rows represent the same customers.

## Generate Match Codes

- Restore your screen view if needed.
- Expand the *Data Quality* section of the *Steps* pane. Double-click the **Match Codes** step to add it to the flow canvas.
- Drag the *Match Codes* step to the right of *CUSTOMERS*. Use your mouse to draw an arrow connecting *Customers* to the *Match Codes* step.

- (Optional) Click  on the flow toolbar to arrange the nodes.



5. In the *Match Codes* tab for the *Match Codes* step, ensure that the default QKB locale, **English (United States)**, is selected.

6. Configure the *Match Codes* step as follows:

- Match column: **Address1**
- New match column name: **[Leave blank]**
- Definition: **Address**
- Sensitivity: **85**



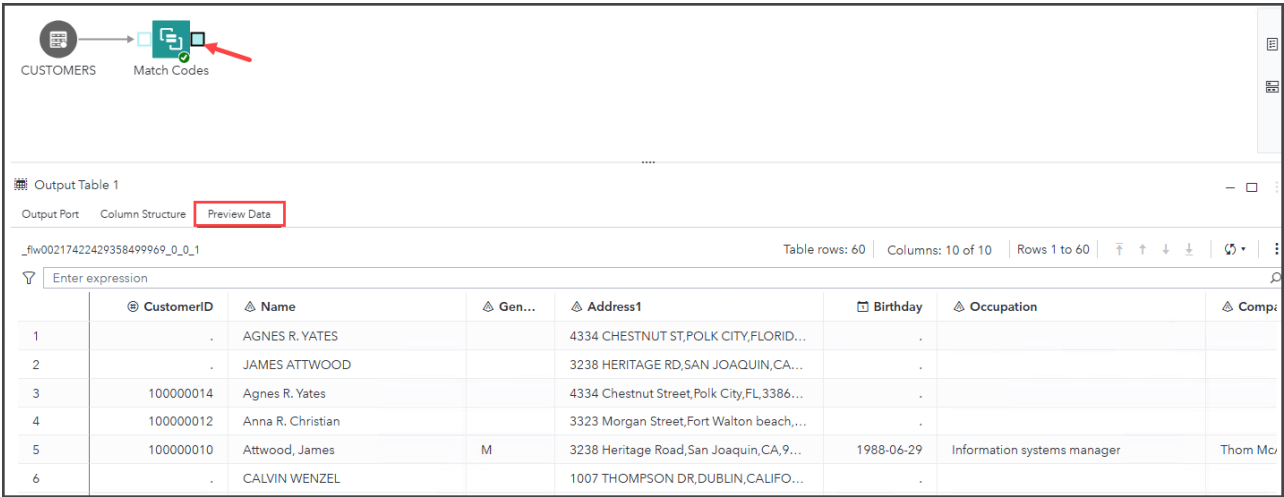
7. Select  to save the flow.

8. Navigate to **SAS Content → Public**. Save the file as **DQ\_HOW.flw**.

9. Select  on the flow toolbar to run the flow.


10. After the flow has run successfully, click the shaded **output port** on *Match Codes*.

11. Click **Preview Data** to review the output.

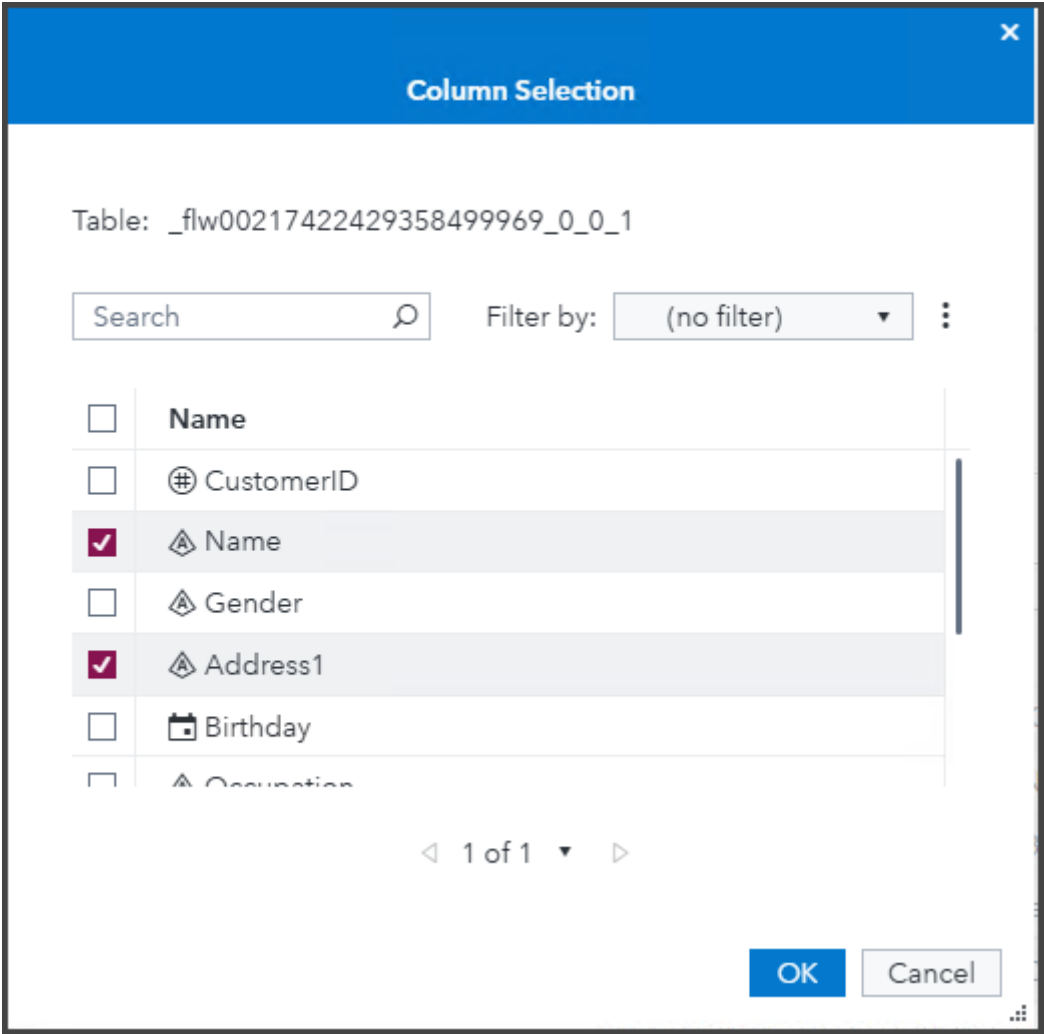


The screenshot shows the SAS Studio interface. At the top, there's a task bar with 'CUSTOMERS' and 'Match Codes'. Below it, the 'Output Table 1' window is open, with the 'Preview Data' tab selected. The table contains 6 rows and 10 columns. The columns are: CustomerID, Name, Gender, Address1, Birthday, Occupation, and Company. The data is as follows:

	CustomerID	Name	Gender	Address1	Birthday	Occupation	Company
1	.	AGNES R. YATES	.	4334 CHESTNUT ST,POLK CITY,FLORID...	.	.	.
2	.	JAMES ATTWOOD	.	3238 HERITAGE RD,SAN JOAQUIN,CA...	.	.	.
3	100000014	Agnes R. Yates	.	4334 Chestnut Street,Polk City,FL,3386...	.	.	.
4	100000012	Anna R. Christian	.	3323 Morgan Street,Fort Walton beach,...	.	.	.
5	100000010	Attwood, James	M	3238 Heritage Road,San Joaquin,CA,9...	1988-06-29	Information systems manager	Thom Mc...
6	.	CALVIN WENZEL	.	1007 THOMPSON DR,DUBLIN,CALIFO...	.	.	.

12. Click  in the *Preview Data* window and select **Column Selection**.

13. In the *Column Selection* window, deselect all columns, then individually select **Name**, **Address1**, and **Address1\_MC85**, then click **OK**.



The screenshot shows the 'Column Selection' dialog box. It has a title bar 'Column Selection' and a close button. Below the title bar, it says 'Table: \_flw00217422429358499969\_0\_0\_1'. There is a search bar and a filter dropdown set to '(no filter)'. A list of columns is shown with checkboxes:

- ☐ Name
- ☐ CustomerID
- ☒ Name
- ☐ Gender
- ☒ Address1
- ☐ Birthday
- ☐ Occupation

At the bottom, there are 'OK' and 'Cancel' buttons. The page number '1 of 1' is also visible.

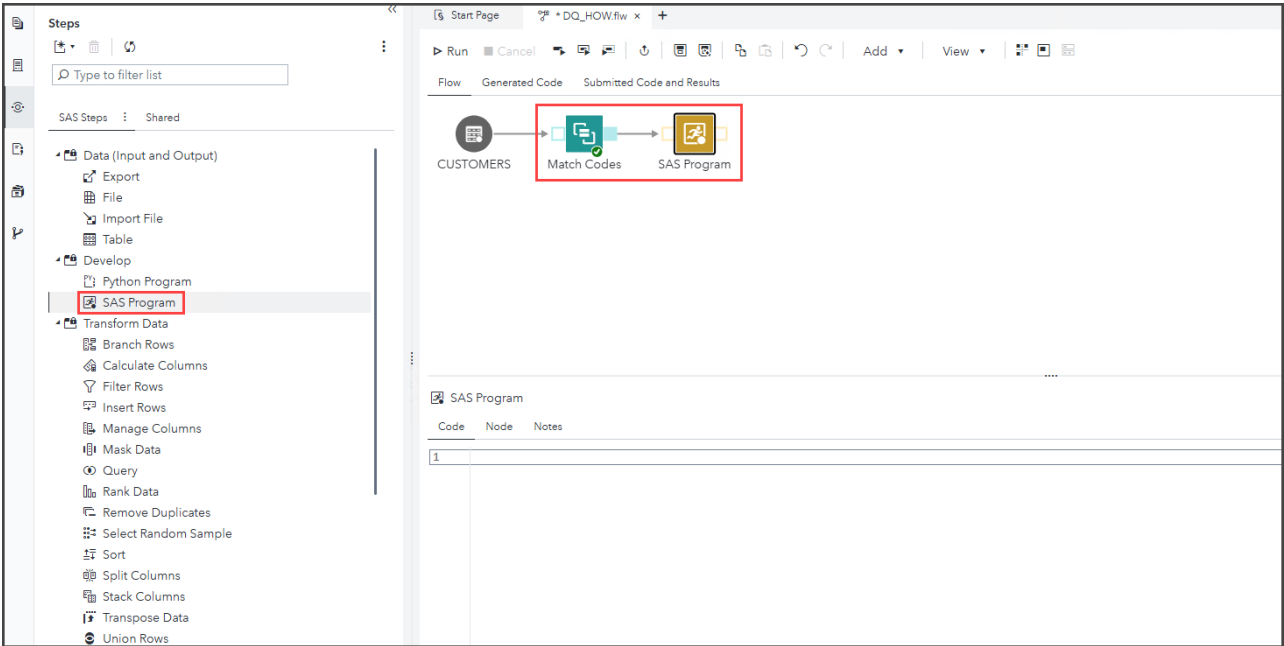
14. Review the output on the *Preview Data* tab.


Output Table 1			
Output Port		Column Structure	Preview Data
_flow0021742242935849969_0_0_1		Table rows: 60   Columns: 3 of 10   Rows 1 to 60	
Enter expression			
	Name	Address1	Address1_MC85
1	AGNES R. YATES	4334 CHESTNUT ST,POLK CITY,FLORIDA,33868,US	\$\$\$\$\$\$\$\$\$\$\$\$\$SKK\$\$\$\$\$\$\$\$\$J2&4-P&-\$\$\$\$\$\$\$\$\$\$\$\$\$...
2	JAMES ATTWOOD	3238 HERITAGE RD,SAN JOAQUIN,CALIFORNIA,93660,US	\$\$\$\$\$\$\$\$\$\$\$\$\$SKHKD\$\$\$\$\$\$\$\$\$2&Y&-&F\$\$\$\$\$\$\$\$\$\$\$\$\$...
3	Agnes R. Yates	4334 Chestnut Street,Polk City,FL,33868,US	\$\$\$\$\$\$\$\$\$\$\$\$\$SKK\$\$\$\$\$\$\$\$\$J2&4-P&-\$\$\$\$\$\$\$\$\$\$\$\$\$...
4	Anna R. Christian	3323 Morgan Street,Fort Walton beach,FL,32548,US	\$\$\$\$\$\$\$\$\$\$\$\$\$SKHK\$\$\$\$\$\$\$\$\$B&YF&P\$\$\$\$\$\$\$\$\$\$\$\$\$...
5	Attwood, James	3238 Heritage Road,San Joaquin,CA,93660,US	\$\$\$\$\$\$\$\$\$\$\$\$\$SKHKD\$\$\$\$\$\$\$\$\$2&Y&-&F\$\$\$\$\$\$\$\$\$\$\$\$\$...
6	CALVIN WENZEL	1007 THOMPSON DR,DUBLIN,CALIFORNIA,94568,US	\$\$\$\$\$\$\$\$\$\$\$\$\$Z00I\$\$\$\$\$\$\$\$\$~2&BN4&P\$\$\$\$\$\$\$\$\$\$\$\$\$...
7	CATHERINE M. THOMAS	3005 BROOKSIDE DRIVE,GUIN,ALABAMA,35563,US	\$\$\$\$\$\$\$\$\$\$\$\$\$SK005\$\$\$\$\$\$\$\$\$MY&34&-\$\$\$\$\$\$\$\$\$\$\$\$\$...
8	CHRISTIAN, ANNA	3323 MORGAN ST,FORT WALTON BEACH,FLORIDA,32548,US	\$\$\$\$\$\$\$\$\$\$\$\$\$SKHK\$\$\$\$\$\$\$\$\$B&YF&P\$\$\$\$\$\$\$\$\$\$\$\$\$...
9	CURRY, KRISTI R.	399 MASONIC HILL ROAD,LITTLE ROCK,ARKANSAS,72212,US	\$\$\$\$\$\$\$\$\$\$\$\$\$SK-\$\$\$\$\$\$\$\$\$B&4&P&32W\$\$\$\$\$\$\$\$\$\$\$\$\$...
10	Calvin B. Wenzel	1007 Thompson drive,Dublin,CA,94568,US	\$\$\$\$\$\$\$\$\$\$\$\$\$Z00I\$\$\$\$\$\$\$\$\$~2&BN4&P\$\$\$\$\$\$\$\$\$\$\$\$\$...
11	Carolan, Kris J.	3935 Lakeland Park Drive,Duluth,GA,30097,US	\$\$\$\$\$\$\$\$\$\$\$\$\$K-K5\$\$\$\$\$\$\$\$\$W&3&W&P-N3\$\$\$\$\$\$\$\$\$\$\$\$\$...
12	Catherine Thomas	3005 Brookside Drive,Guin,AL,35563,US	\$\$\$\$\$\$\$\$\$\$\$\$\$SK005\$\$\$\$\$\$\$\$\$MY&34&-\$\$\$\$\$\$\$\$\$\$\$\$\$...
13	Cheeks, Linda	3385 Rinehart Road,Miami,FL,33128,US	\$\$\$\$\$\$\$\$\$\$\$\$\$SKKD\$\$\$\$\$\$\$\$\$Y&P&Y-\$\$\$\$\$\$\$\$\$\$\$\$\$...
14	Christian, Anna R.	3323 Morgan Street,Fort Walton beach,FL,32548,US	\$\$\$\$\$\$\$\$\$\$\$\$\$SKHK\$\$\$\$\$\$\$\$\$B&YF&P\$\$\$\$\$\$\$\$\$\$\$\$\$...

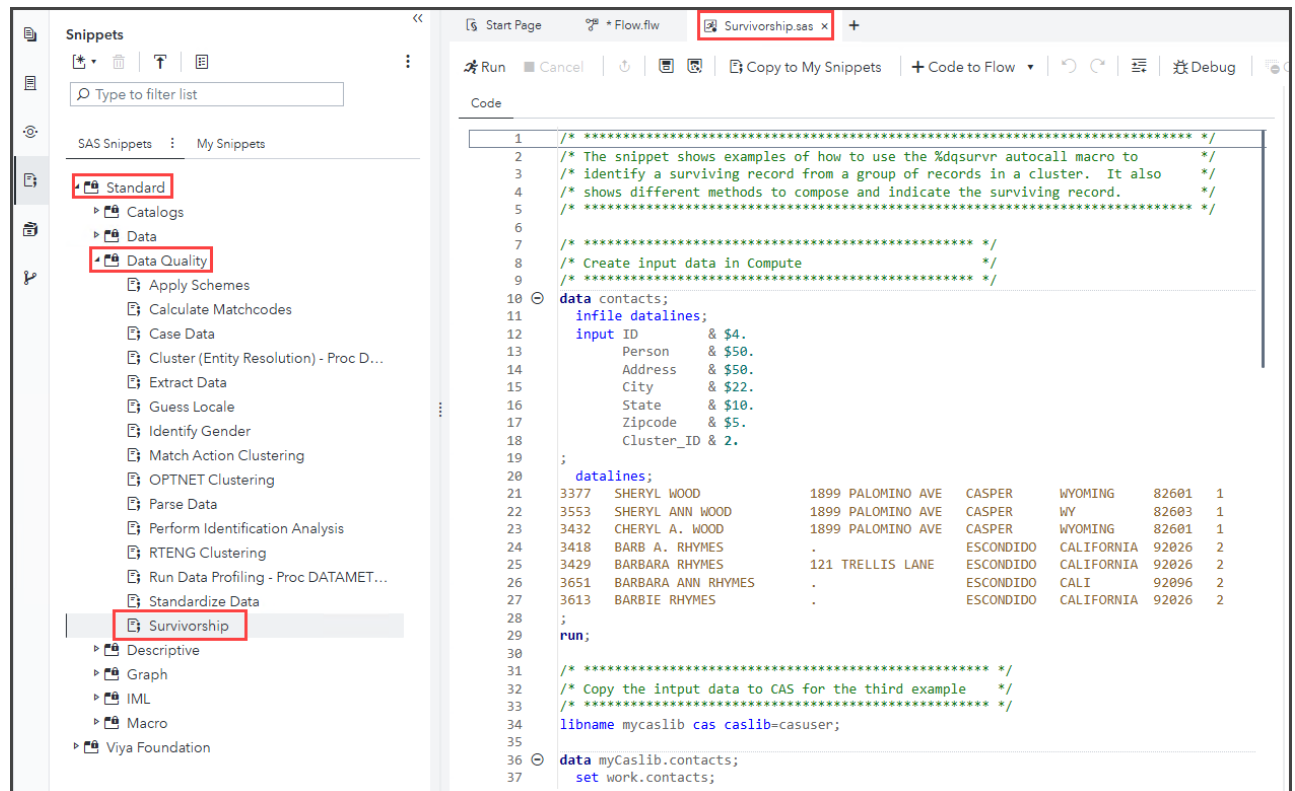
Notice that rows with equivalent *Address1* values generated the same match codes regardless of *Address1* format. We can use these match codes to combine customer data into one surviving record.

Remove Duplicates with Survivorship Snippet

1. Add the **SAS Program** step from the *Develop* section of the *Steps* pane to the flow canvas and connect it to the *Match Nodes* step.
2. **Arrange** the nodes and **save** the flow.



3. Select  to open the **Snippets** pane.
4. Expand **Standard** → **Data Quality** and double-click **Survivorship** to open the *Survivorship* snippet in a new tab.



Per the snippet documentation: *The snippet shows examples of how to use the %dqsurvr autocall macro to identify a surviving record from a group of records in a cluster. It also shows different methods to compose and indicate the surviving record.*

- Copy the first example that calls the %dqsurvr macro (**lines 52 - 59**) and paste it into the SAS Program node's Code tab.

#### ► Details

Click to view or copy the required section of code.

```

%dqsurvr (inTable=contacts,
          outTable=work.contacts_out,
          clusterColumn=cluster_id,
          rowRule1=(max,ID),
          firstColumnRule1=(highocc,State),
          firstColumnRuleAppliedCols=(Zipcode),
          secondColumnRule1=(not_missing, Address),
          keepDuplicates=0);

```

```

1 %dqsurv (inTable=contacts,
2         outTable=work.contacts_out,
3         clusterColumn=cluster_id,
4         rowRule1=(max, ID),
5         firstColumnRule1=(highocc, State),
6         firstColumnRuleAppliedCols=(Zipcode),
7         secondColumnRule1=(not_missing, Address),
8         keepDuplicates=0);
9

```

6. Edit the macro parameters as follows:

- inTable=**&\_input1**

✎ **&\_input1** is a default macro variable which stores the input table name for a node. This option is used because the input table is a temporary table created by the Match Codes step, and temporary tables are given a random name by default.

- outTable=**&\_output1**

✎ **&\_output1** is a macro variable which stores the output table name for a node. This option is used because we want to store this output in a temporary table generated by the SAS Program step, and temporary tables are given a random name by default.

- clusterColumn=**Address1\_MC85**

- rowRule1=**(longest, Name)**

✎ The surviving record will be chosen by the longest **Name** value, because this should be the fullest version of the customer's name.

- firstColumnRule1=**(not\_missing, Gender)**

✎ The surviving record should store the first non-missing *Gender* value from the cluster records.

- firstColumnRuleAppliedCols=**(Birthday, Occupation, Company)**

✎ The surviving record should store the *Birthday*, *Occupation*, and *Company* values from the row satisfying *firstColumnRule1*.

- secondColumnRule1=**(not\_missing, Email)**

✎ The surviving record should store the first non-missing *Email* value from the cluster records.

- **Add secondColumnRuleAppliedCols=(CustomerID, Phone)**, (including comma after option)

✎ The surviving record should store the *CustomerID* and *Phone* values from the row satisfying *secondColumnRule1*.



- **Add a comma after** keepDuplicates=0
- **Add generateDistinctSurvivor=1** after keepDuplicates=0

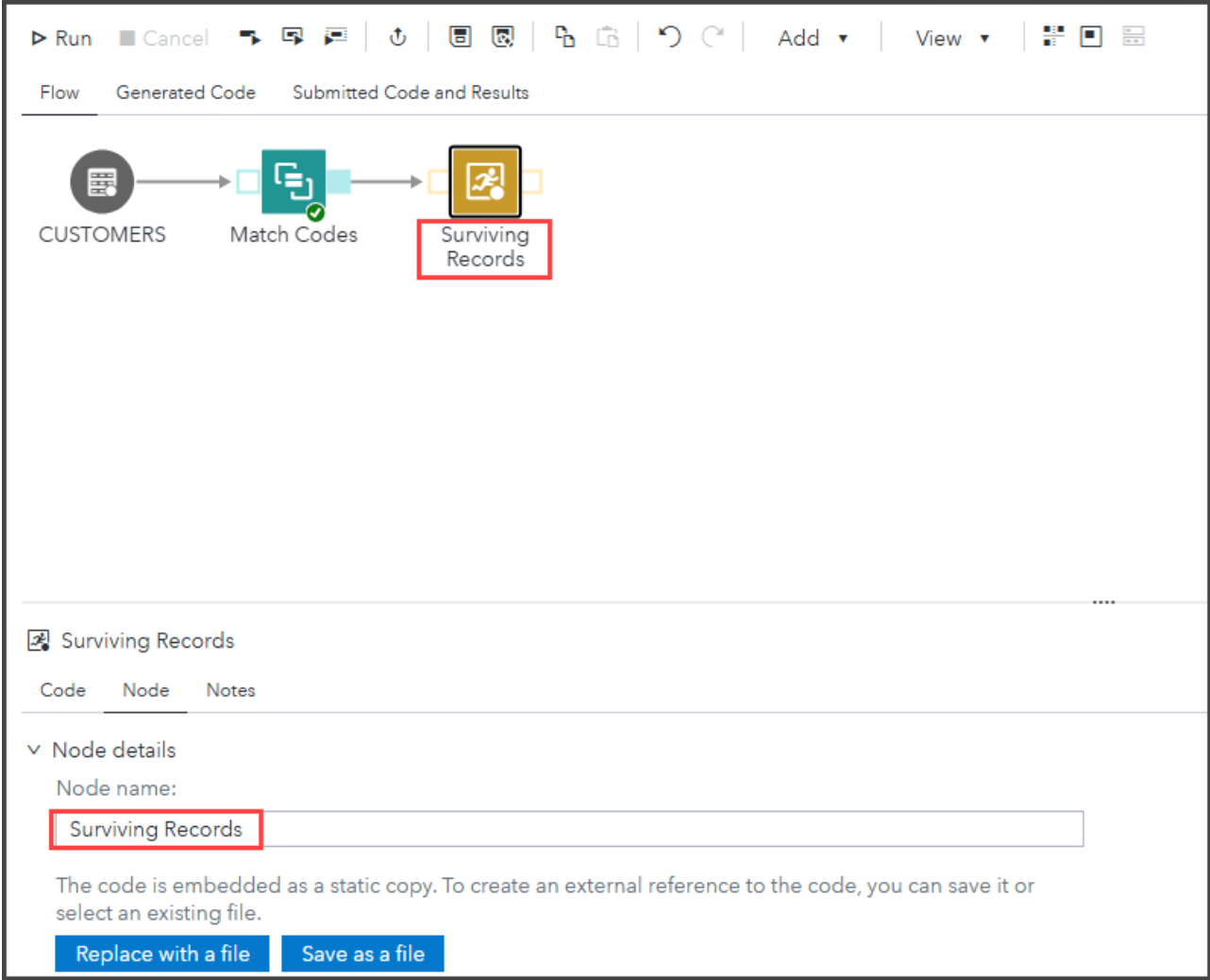
► Details

Click to view or copy the edited macro call.

```
%dqsrvr (inTable=&_input1,
        outTable=&_output1,
        clusterColumn=Address1_MC85,
        rowRule1=(longest,Name),
        firstColumnRule1=(not_missing,Gender),
        firstColumnRuleAppliedCols=(Birthday,Occupation,Company),
        secondColumnRule1=(not_missing,Email),
        secondColumnRuleAppliedCols=(CustomerID,Phone),
        keepDuplicates=0,
        generateDistinctSurvivor=1);
```

SAS Program		
Code	Node	Notes
1	%dqsrvr	(inTable=&_input1,
2		outTable=&_output1,
3		clusterColumn=Address1_MC85,
4		rowRule1=(longest,Name),
5		firstColumnRule1=(not_missing,Gender),
6		firstColumnRuleAppliedCols=(Birthday,Occupation,Company),
7		secondColumnRule1=(not_missing,Email),
8		secondColumnRuleAppliedCols=(CustomerID,Phone),
9		keepDuplicates=0,
10		generateDistinctSurvivor=1);

7. On the **Node** tab, re-name the node to **Surviving Records**.



- 8. **Save** the changes to the flow.
- 9. **Right-click** *Surviving Records* and select **Run node**. Select the **output port** of *Surviving Records* and **preview** the results.

Output table 1

Output Port   Column Structure   Preview Data

\_flw002174224494166611006\_0\_0\_1


Table rows: 20 | Columns: 10 of 10 | Rows 1 to 20 | [Filter] [Sort] [Refresh] [More]

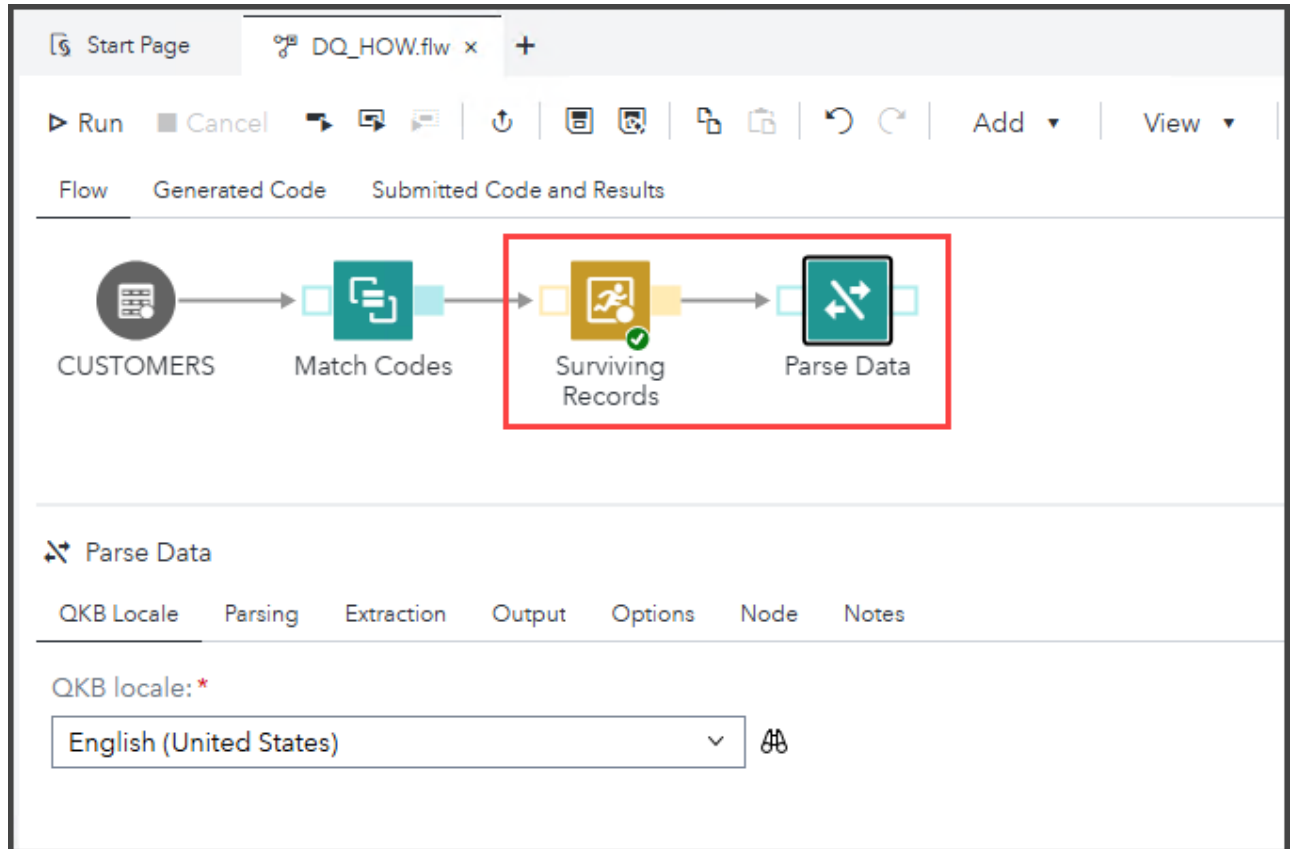
Enter expression

	CustomerID	Name	Gen...	Address1	Birthday	Occupation	Comp:
1	100000016	NICHOLE W. ROBINSON	F	622 CHESTNUT ST,TAMPA,FLORIDA,33...	1999-07-05	Architectural drafter	A Plus Lav
2	100000019	Jenkins, Thomas E.	M	2565 Pin Oak Drive,Davenport,IA,5280...	1972-10-26	Certified pest control applicator	Gemco
3	100000003	Samuel C. Tennyson	M	2715 Mulberry avenue,Donaldson,AR,7...	1978-01-27	Land agent	AM/PM C.
4	100000005	Smith, Wanda R.	F	2190 Cedar Street,Pine Bluff,AR,71601,US	1999-01-26	Insurance sales agent	Quality M
5	100000004	CURRY, KRISTI R.	F	399 MASONIC HILL ROAD,LITTLE ROC...	1996-05-13	Industrial property manager	Giant
6	100000018	Carolann, Kris J.	M	3935 Lakeland Park Driv,Duluth,GA,300...	1986-09-23	Real property appraiser	Dun Rite L
7	100000001	CATHERINE M. THOMAS	F	3005 BROOKSIDE DRIVE,GUIN,ALABA...	1983-06-12	Heating air-conditioning and refrigerati...	House of i
8	100000015	REGINA C. SNIDER	F	3697 RINEHART RD,SUNRISE,FLORIDA,...	1989-09-22	Automotive glass installer	Fradkin Br
9	100000010	Attwood, James	M	3238 Heritage Road,San Joaquin,CA,9...	1988-06-29	Information systems manager	Thom Mc
10	100000011	Kaye, Echo P.	F	3215 Thompson Drive,San Leandro,CA,...	1966-06-12	Navy	Canal Ville

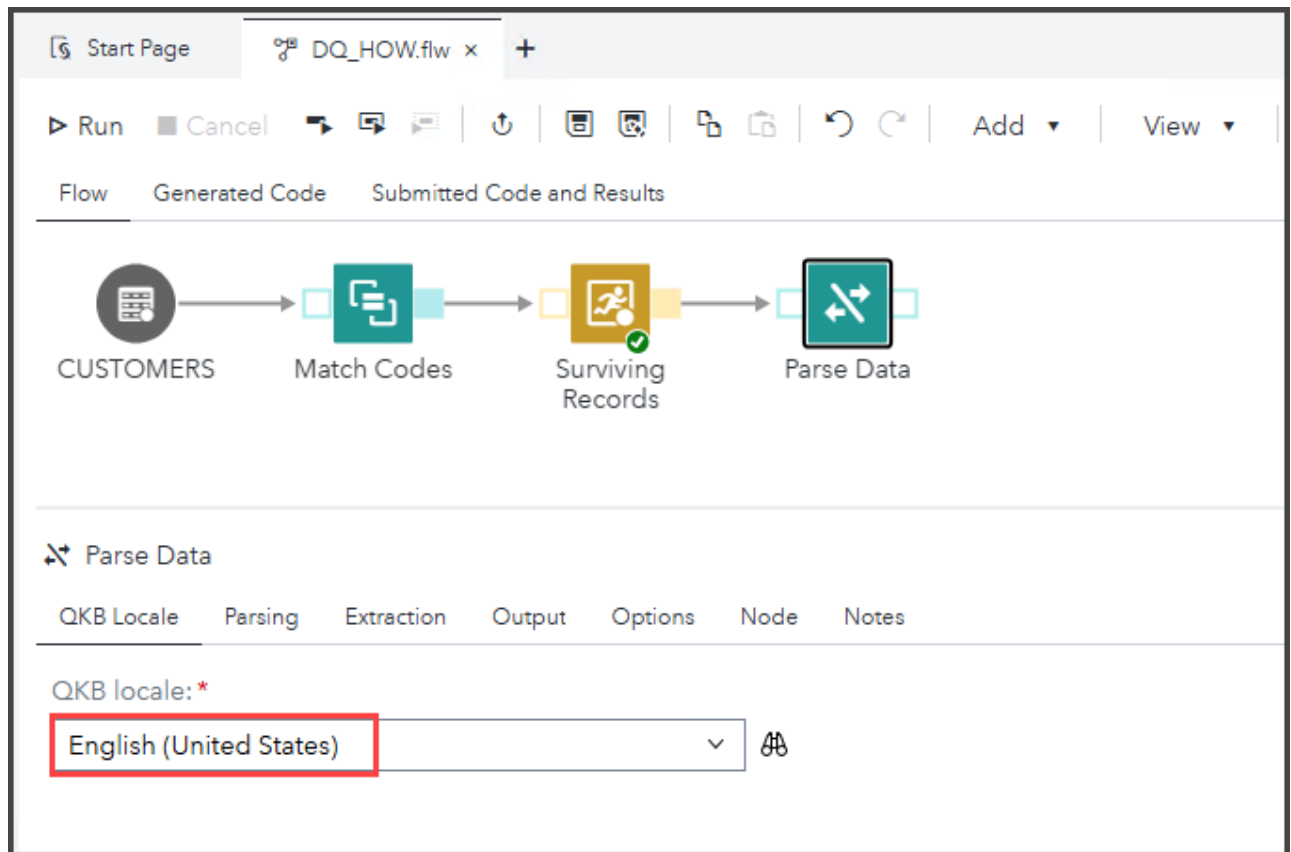
The output table has **20** rows and **5** columns. Duplicate records were accurately condensed to store all customer data in one record. Next, we'll parse *Address1* and store individual tokens like Street, City, and State in separate columns.

Parse Address Data

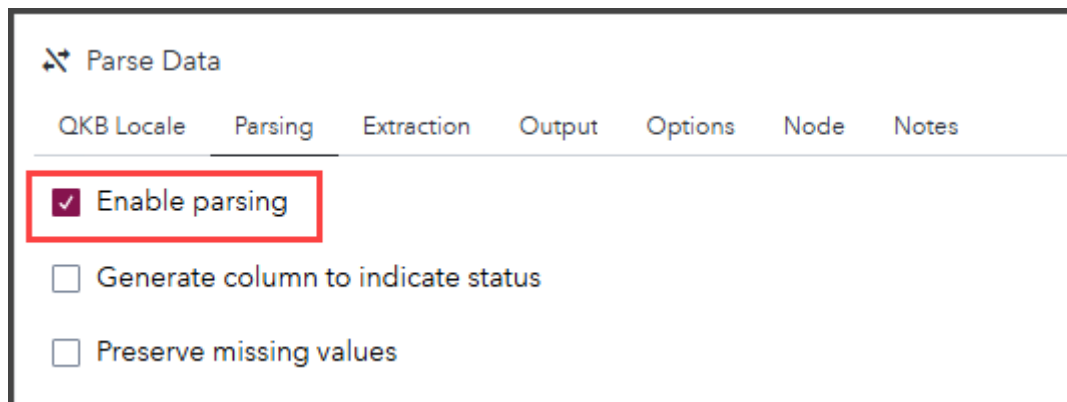
1. **(REQUIRED)** Select  to return to the **Steps** pane.
2. Add the **Parse Data** step from the *Data Quality* section of the *Steps* pane to the flow canvas and connect it to the *SAS Program (Surviving Records)* step.
3. **Arrange** the nodes and **save** the flow.



4. In the *QKB Locale* tab for the *Parse Data* step, ensure that the default QKB locale, **English (United States)**, is selected.

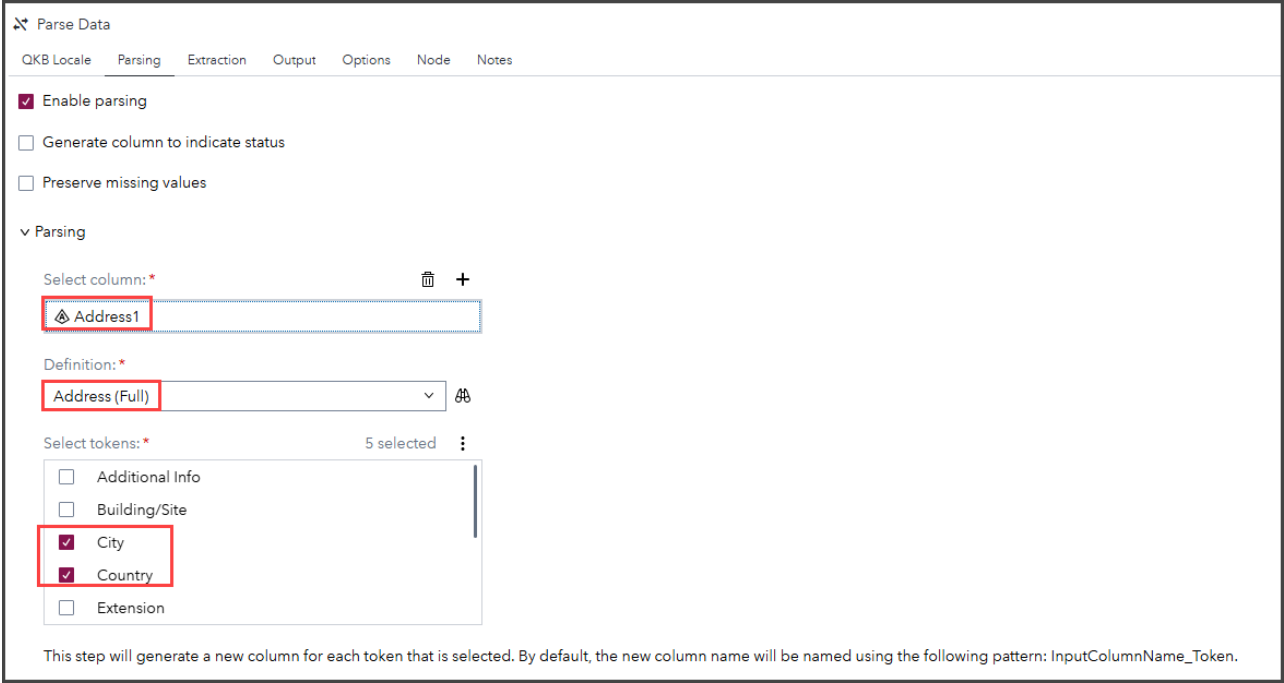


5. On the *Parsing* tab, check the box to **Enable parsing**.



6. Configure the *Parse Data* step as follows:

- Select column: **Address1**
- Definition: **Address (Full)**
- Select tokens: **City, Country, Postal Code, State/Province, Street**



7. **Save** the changes to the flow.
8. **Right-click** *Parse Data* and select **Run node**. Select the **output port** of *Parse Data* and **preview** the results.

Use the *Column Selection* option or scroll all the way to the right in the *Preview Data* window to see new columns.

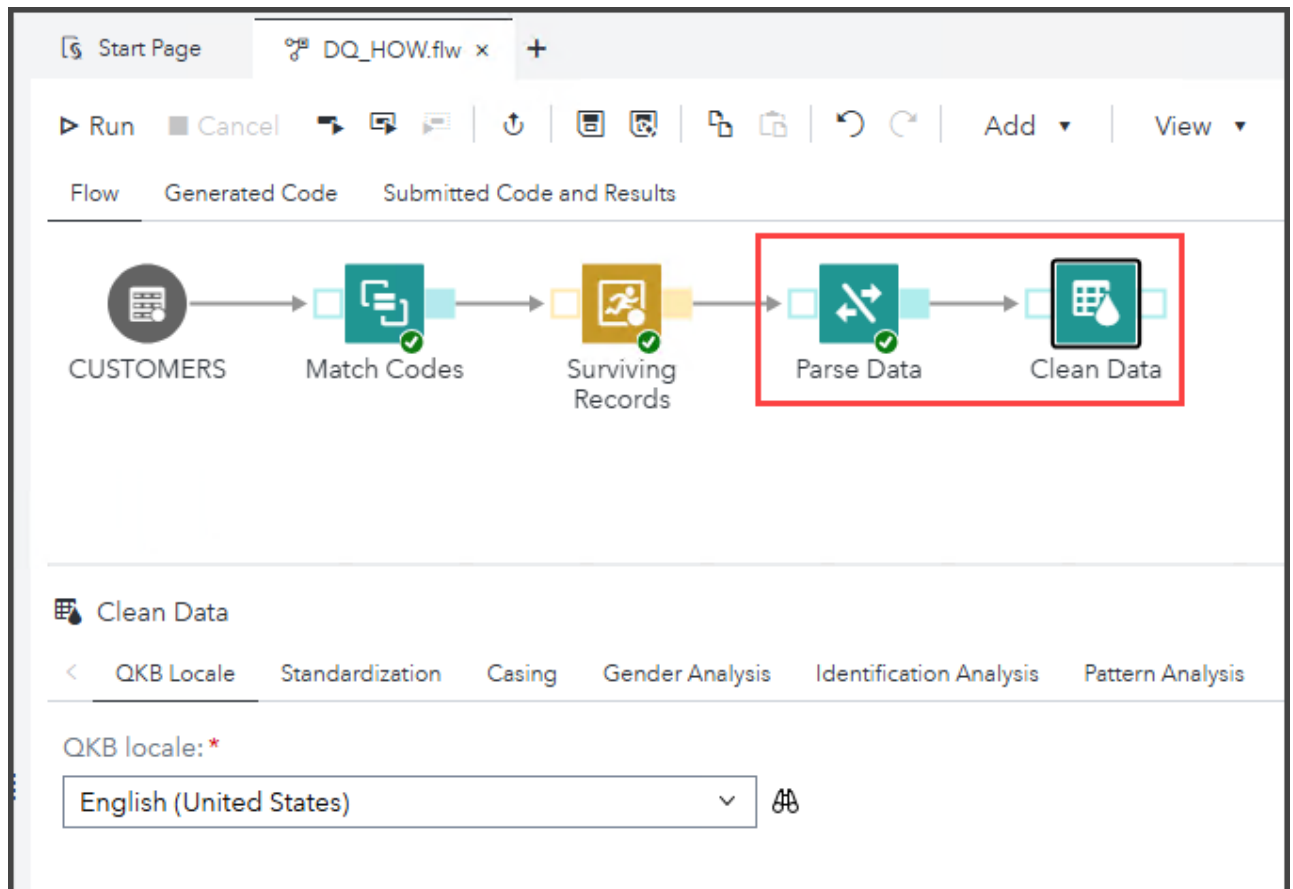
Output Table 1						
_flow001174224653426512730_0_0_3						
Table rows: 20   Columns: 6 of 15   Rows 1 to 20						
Enter expression						
	Address1	Address1_CITY	Address1_PO...	Address1_STATE_...	Address1_STREET	Address1_C...
1	622 CHESTNUT ST,TAMPA,FLORIDA,33610,US	TAMPA	33610	FLORIDA	622 CHESTNUT ST	US
2	2565 Pin Oak Drive,Davenport,IA,52803,US	Davenport	52803	IA	2565 Pin Oak Drive	US
3	2715 Mulberry avenue,Donaldson,AR,71941,US	Donaldson	71941	AR	2715 Mulberry avenue	US
4	2190 Cedar Street,Pine Bluff,AR,71601,US	Pine Bluff	71601	AR	2190 Cedar Street	US
5	399 MASONIC HILL ROAD,LITTLE ROCK,ARKANSAS,72212,US	LITTLE ROCK	72212	ARKANSAS	399 MASONIC HILL ROAD	US
6	3935 Lakeland Park Driv,Duluth,GA,30097,US	Duluth	30097	GA	3935 Lakeland Park Driv	US
7	3005 BROOKSIDE DRIVE,GUIN,ALABAMA,35563,US	GUIN	35563	ALABAMA	3005 BROOKSIDE DRIVE	US
8	3697 RINEHART RD,SUNRISE,FLORIDA,33323,US	SUNRISE	33323	FLORIDA	3697 RINEHART RD	US
9	3238 Heritage Road,San Joaquin,CA,93660,US	San Joaquin	93660	CA	3238 Heritage Road	US
10	3215 Thompson Drive,San Leandro,CA,94578,US	San Leandro	94578	CA	3215 Thompson Drive	US

The output table has **5** new columns storing the **City**, **Country**, **Postal Code**, **State/Province**, and **Street** from each *Address1* value. Note that:

- **Address1\_CITY** varies in casing. Some city names are written in uppercase, while others are in proper case.
- **Address1\_STATE\_PROVINCE** varies in format. Some state names are written in full, while others are written in abbreviations.
- **Address1\_STREET** varies in casing and format. Specifically, some street addresses use full road names, while others use abbreviations.
- **Address1\_COUNTRY** uses a consistent country abbreviation standard, but we'd like to see the full country name instead. We'll correct these issues and others with the Clean Data step.

Standardize Data

1. Add the **Clean Data** step from the *Data Quality* section of the *Steps* pane to the flow canvas and connect it to the *Parse Data* step.
2. **Arrange** the nodes and **save** the flow.



3. In the *QKB Locale* tab for the *Clean Data* step, ensure that the default QKB locale, **English (United States)**, is selected.
4. On the *Standardization* tab, check the box to **Enable standardization**.
5. Configure the *Clean Data* step as follows:
  - Select column: **Name**
  - Definition: **Name**
  - Column options: **Create new column**
  - New column: **[Leave blank]**

Clean Data

<

QKB Locale

Standardization

Casing

Gender Analysis

Identification Analysis

Pattern Analysis

☒

Enable standardization

☐

Generate column to indicate status

☐

Preserve missing values

▼

Standardize

You can perform up to 10 standardization operations.

Select column: \*

+

Name

Definition: \*

Name

▼

Column options:

☐ Replace existing column

☒ Create new column

New column:

If no value is provided, the new column name is in the format InputColumnName\_STND.

6. Under the *Additional Standardize* heading, repeat this process to standardize columns as specified below. **For all *Additional Standardize* columns, select *Create new column* and leave the new column name *blank*.**

Select Column	Definition
Address1_CITY	City
Address1_STATE_PROVINCE	State/Province (Full Name)
Address1_STREET	Address
Address1_COUNTRY	Country

7. **Save** the changes to the flow.

8. **Right-click** *Clean Data* and select **Run node**. Select the **output port** of *Clean Data* and **preview** the results.

15 / 20

✖ If the *Clean Data* step fails with *ERROR: Key not found* and *ERROR: Error in the FILENAME statement*, simply run the step again. This is a known bug which is resolved by executing again.

Flow

Generated Code

Submitted Code and Results

Code

Log

Output Data (1)

⊗ Errors (2)

⚠ Warnings (0)

ℹ Notes (32)

⊗ ERROR: Key not found.

⊗ ERROR: Error in the FILENAME statement.

9. (Optional) use the *Column selection* option to limit the column view in the *Preview Data* window.

Output Table 1

Output Port

Column Structure

Preview Data

\_fiw00117423203662251754\_0\_0\_1

Table rows: 20 | Columns: 2 of 20 | Rows 1 to 20

Enter expression

	Name	Name_STND
1	NICHOLE W. ROBINSON	Nichole W Robinson
2	Jenkins, Thomas E.	Thomas E Jenkins
3	Samuel C. Tennyson	Samuel C Tennyson
4	Smith, Wanda R.	Wanda R Smith
5	CURRY, KRISTI R.	Kristi R Curry
6	Carolan, Kris J.	Kris J Carolan
7	CATHERINE M. THOMAS	Catherine M Thomas
8	REGINA C. SNIDER	Regina C Snider
9	Attwood, James	James Attwood
10	Kaye, Echo P.	Echo P Kaye

Standardized names are proper cased and written in **First MI Last** format.

Output Table 1

Output Port

Column Structure

Preview Data

\_fiw00117423203662251754\_0\_0\_1

Table rows: 20 | Columns: 6 of 20 | Rows 1 to 20

Enter expression

	Address1_CITY	Address1_STATE_...	Address1_COUNTRY	Address1_CITY_STND	Address1_STATE_...	Address1_COUNTRY_STND
1	TAMPA	FLORIDA	US	Tampa	Florida	United States of America
2	Davenport	IA	US	Davenport	Iowa	United States of America
3	Donaldson	AR	US	Donaldson	Arkansas	United States of America
4	Pine Bluff	AR	US	Pine Bluff	Arkansas	United States of America
5	LITTLE ROCK	ARKANSAS	US	Little Rock	Arkansas	United States of America
6	Duluth	GA	US	Duluth	Georgia	United States of America
7	GUIN	ALABAMA	US	Guin	Alabama	United States of America
8	SUNRISE	FLORIDA	US	Sunrise	Florida	United States of America
9	San Joaquin	CA	US	San Joaquin	California	United States of America
10	San Leandro	CA	US	San Leandro	California	United States of America

Standardized city, state, and country names are proper cased and written in full instead of abbreviated.



Output Table 1

Output PortColumn StructurePreview Data

\_flw00117423203662251754\_0\_0\_1

Table rows: 20Columns: 2 of 20Rows 1 to 20

Enter expression

	Address1_STREET	Address1_STREET_STND
1	622 CHESTNUT ST	622 Chestnut St
2	2565 Pin Oak Drive	2565 Pin Oak Dr
3	2715 Mulberry avenue	2715 Mulberry Ave
4	2190 Cedar Street	2190 Cedar St
5	399 MASONIC HILL ROAD	399 Masonic Hill Rd
6	3935 Lakeland Park Drive	3935 Lakeland Park Dr
7	3005 BROOKSIDE DRIVE	3005 Brookside Dr
8	3697 RINEHART RD	3697 Rinehart Rd
9	3238 Heritage Road	3238 Heritage Rd
10	3215 Thompson Drive	3215 Thompson Dr

Standardized street names are proper cased and road type names (Street, Drive, Avenue, etc) are shortened.

We'll finish by creating an output table to store the final results.

Create Output Table

- 1. Add the **Query** step (from the *Transform Data* section of the *Steps* pane) to the flow canvas and connect it to the *Clean Data* step.
- 2. Right-click the *Query* step output port and select **Add a table**.
- 3. **Arrange** the nodes and **save** the flow.

Start Page

\* DQ\_HOW.flw x +

RunCancel

FlowGenerated CodeSubmitted Code and Results

CUSTOMERSMatch CodesSurviving RecordsParse DataClean DataQueryTable

Query

OptionsNodeNotes

Columns

+ Calculated ColumnFilter

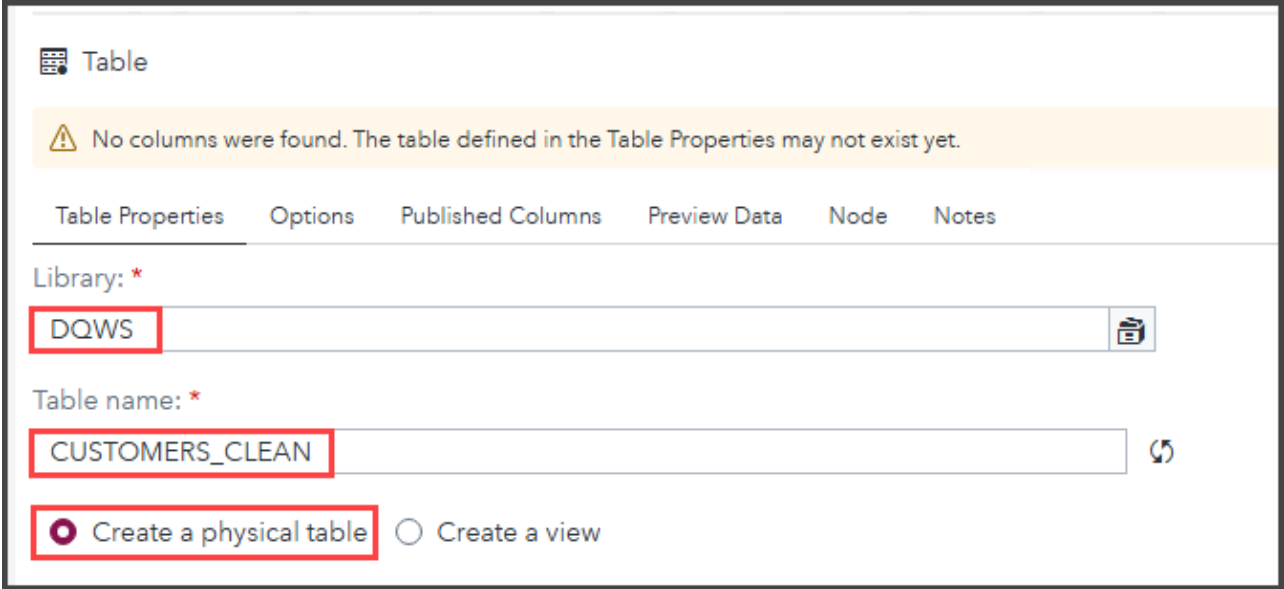
t1 (Clean Data)

SelectJoinFilterSortOutput Options

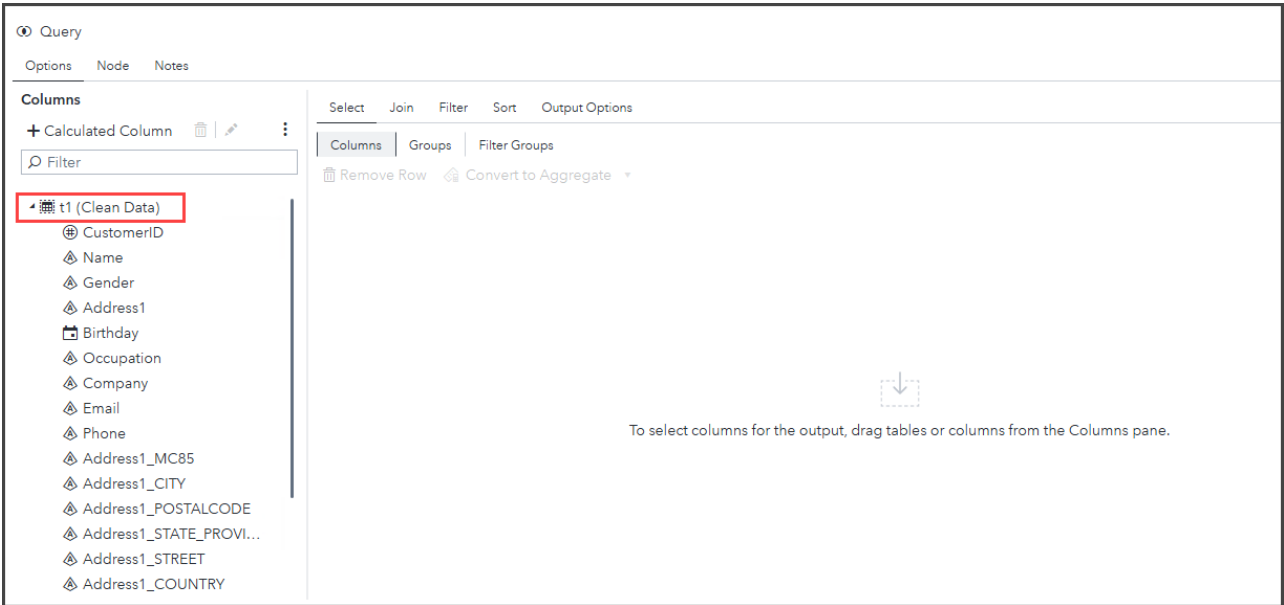
ColumnsGroupsFilter Groups

Remove RowConvert to Aggregate

4. Select the **Table** node on the flow canvas. In the **Table Properties** section, type the following table attributes:
- Library: **DQWS**
  - Table name: **CUSTOMERS\_CLEAN**
  - Select **Create a physical table** (default selection)



5. Select the **Query** node on the flow canvas. In the *Options* tab, expand **t1 (Clean Data)** to view all source columns.



6. Add the following columns to the *Select* tab by double-clicking the column name on the left. Update the column attributes as specified below:

Source	Name	Label
CustomerID	CustomerID	<b>Customer ID</b>
Name_STND	<b>Name</b>	
Address1_STREET_STND	<b>Street</b>	

Source	Name	Label
Address1_CITY_STND	<b>City</b>	
Address1_STATE_PROVINCE_STND	<b>State</b>	
Address1_POSTALCODE	<b>PostalCode</b>	<b>Postal Code</b>
Address1_COUNTRY_STND	<b>Country</b>	
Gender	Gender	
Birthday	Birthday	
Occupation	Occupation	
Company	Company	
Email	Email	
Phone	Phone	

✎ Updates to column attributes are **bolded**.

Query

OptionsNodeNotes

Columns

+ Calculated Column

Filter

t1 (Clean Data)

CustomerID

Name

Gender

Address1

Birthday

Occupation

Company

Email

Phone

Address1\_MC85

Address1\_CITY

Address1\_POSTALCODE

Address1\_STATE\_PROVINCE

Address1\_STREET

Address1\_COUNTRY

Name\_STND

Address1\_CITY\_STND

Address1\_STATE\_PROVIN...

Address1\_STREET\_STND

Address1\_COUNTRY\_STND



SelectJoinFilterSortOutput Options

ColumnsGroupsFilter Groups

Remove RowConvert to Aggregate

	Table	Source	Name	Label	Type	Length	Format
<input type="checkbox"/>	t1	CustomerID	CustomerID	Customer ID	Numeric	8	BEST12.
<input type="checkbox"/>	t1	Name_STND	Name		Character	255	\$255.
<input type="checkbox"/>	t1	Address1_STRE...	Street		Character	255	\$255.
<input type="checkbox"/>	t1	Address1_CITY...	City		Character	255	\$255.
<input type="checkbox"/>	t1	Address1_STAT...	State		Character	255	\$255.
<input type="checkbox"/>	t1	Address1_POST...	PostalCode	Postal Code	Character	200	
<input type="checkbox"/>	t1	Address1_COU...	Country		Character	255	\$255.
<input type="checkbox"/>	t1	Gender	Gender		Character	1	\$1.
<input type="checkbox"/>	t1	Birthday	Birthday		Numeric	8	YYMMD...
<input type="checkbox"/>	t1	Occupation	Occupation		Character	46	\$46.
<input type="checkbox"/>	t1	Company	Company		Character	27	\$27.
<input type="checkbox"/>	t1	Email	Email		Character	30	\$30.
<input type="checkbox"/>	t1	Phone	Phone		Character	12	\$12.

7. Click the *Sort* tab.
8. Double-click *CustomerID* to add it to the *Sort* tab. Keep the default sort order selection (**Ascending**).
9. **Save** the changes to the flow.
10. **Right-click** *Query* and select **Run node**. Select the output table **CUSTOMERS\_CLEAN** and **preview** the results.

✎ If needed, click  to hide the Steps pane, or click  to maximize the preview.

CUSTOMERS\_CLEAN

Table PropertiesOptionsPublished ColumnsPreview DataNodeNotes

CUSTOMERS\_CLEAN

Table rows: 20Columns: 13 of 13Rows 1 to 20

Enter expression

	CustomerID	Name	Street	City	State	PostalCode	Country	G..	Birthday	Occupation
1	100000001	Catherine M Thomas	3005 Brookside Dr	Guin	Alabama	35563	United States of America	F	1983-06-12	Heating air-conditioning and refrigerati...
2	100000002	John L von Neumann	4026 Maple Ln	Huntsville	Alabama	35802	United States of America	M	1995-03-04	Farm equipment mechanic
3	100000003	Samuel C Tennyson	2715 Mulberry Ave	Donaldson	Arkansas	71941	United States of America	M	1978-01-27	Land agent
4	100000004	Kristi R Curry	399 Masonic Hill Rd	Little Rock	Arkansas	72212	United States of America	F	1996-05-13	Industrial property manager
5	100000005	Wanda R Smith	2190 Cedar St	Pine Bluff	Arkansas	71601	United States of America	F	1999-01-26	Insurance sales agent
6	100000006	James R R Hoadley	1078 Davis Ave	Concord	California	94520	United States of America	M	1995-06-10	Marine electronics technician
7	100000007	Calvin B Wenzel	1007 Thompson Dr	Dublin	California	94568	United States of America	M	1966-09-28	Medical social worker
8	100000008	Jennifer R Ryan	4974 Water St	Fremont	California	94539	United States of America	F	1970-12-15	Diabetes management nurse
9	100000009	Richard T Reyes	4819 Kerry Way	Los Angeles	California	90017	United States of America	M	1974-10-06	Forest fire inspector
10	100000010	James Attwood	3238 Heritage Rd	San Joaquin	California	93660	United States of America	M	1988-06-29	Information systems manager
11	100000011	Echo P Kaye	3215 Thompson Dr	San Leandro	California	94578	United States of America	F	1966-06-12	Navy
12	100000012	Anna R Christian	3323 Morgan St	Fort Walton Beach	Florida	32548	United States of America	F	1979-07-30	Mixing and blending machine tender
13	100000013	Linda R Cheeks	3385 Rinehart Rd	Miami	Florida	33128	United States of America	F	1999-09-30	Delivery service driver
14	100000014	Agnes R Yates	4334 Chestnut St	Polk City	Florida	33868	United States of America	F	1998-05-13	Floor installer
15	100000015	Regina C Snider	3697 Rinehart Rd	Sunrise	Florida	33323	United States of America	F	1989-09-22	Automotive glass installer
16	100000016	Nichole W Robinson	622 Chestnut St	Tampa	Florida	33610	United States of America	F	1999-07-05	Architectural drafter

11. **Close** the *DQ\_HOW.flw* file tab after reviewing the results.

## Exercise Completed

**You have completed the exercise on building a flow with SAS Data Quality!**

**THANK YOU FOR ATTENDING THIS WORKSHOP!**