

## Be Prepared: An Introduction to SAS® Data Preparation

Mary Kathryn Queen, SAS Institute Inc., Cary, NC

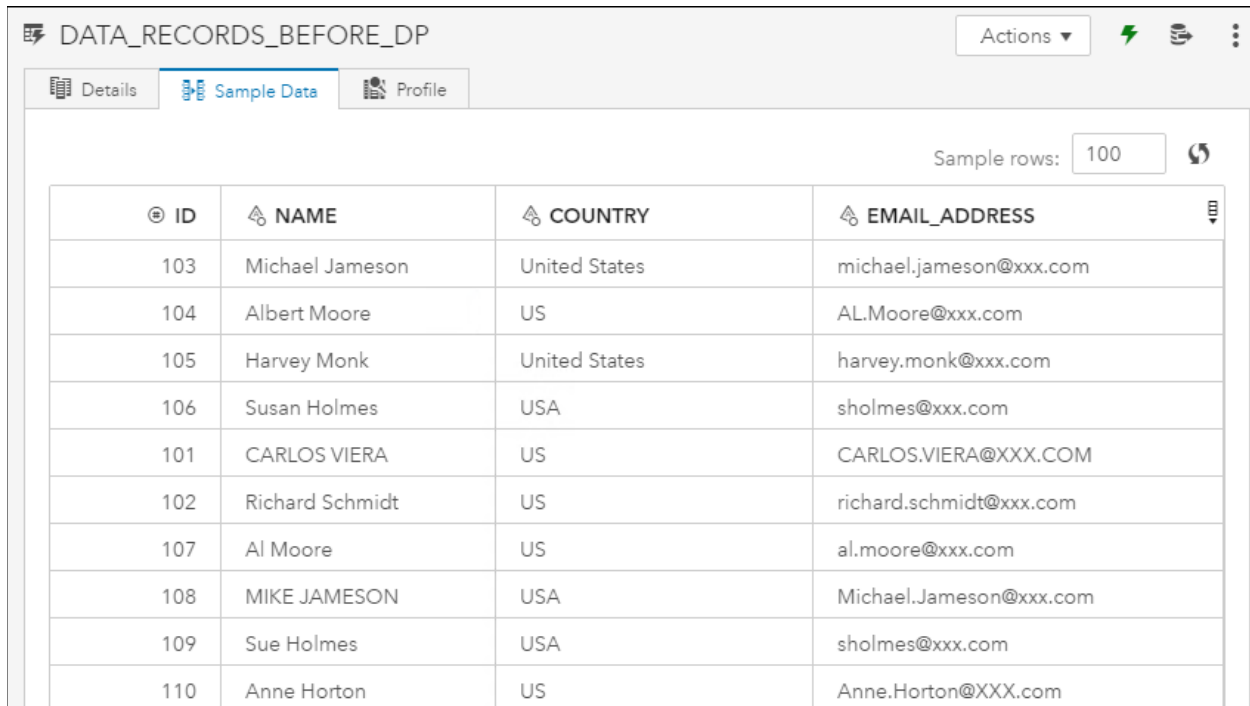
### ABSTRACT

"Be Prepared" is not just a motto for the Boy Scouts of America; it is also an important concept for your data, on which you base your business decisions. SAS® Data Preparation powered by SAS® Viya® provides self-service capabilities for preparing your data to create more consistent and accurate reports or analytic models, which ultimately lead to better and more informed business decisions. This presentation walks through using the profiling and data cleansing features of SAS Data Preparation to show how a non-technical person can use the point-and-click interface to prepare their data.

### INTRODUCTION

Adding a SAS Data Preparation license to your SAS Viya installation gives you the capability to perform advanced profiling and data discovery in SAS Data Explorer. Using these results you can create your data preparation plan, which transforms your data in SAS Data Studio to suit your report or analytical model needs.

The data set in Figure 1 is used to illustrate the self-service capabilities of SAS Data Preparation in SAS Viya:



DATA\_RECORDS\_BEFORE\_DP

Actions

Details Sample Data Profile

Sample rows: 100

ID	NAME	COUNTRY	EMAIL_ADDRESS
103	Michael Jameson	United States	michael.jameson@xxx.com
104	Albert Moore	US	AL.Moore@xxx.com
105	Harvey Monk	United States	harvey.monk@xxx.com
106	Susan Holmes	USA	sholmes@xxx.com
101	CARLOS VIERA	US	CARLOS.VIERA@XXX.COM
102	Richard Schmidt	US	richard.schmidt@xxx.com
107	Al Moore	US	al.moore@xxx.com
108	MIKE JAMESON	USA	Michael.Jameson@xxx.com
109	Sue Holmes	USA	sholmes@xxx.com
110	Anne Horton	US	Anne.Horton@XXX.com

Figure 1. Data Records before Data Preparation

## ADVANCED PROFILING AND DATA DISCOVERY

With SAS Data Preparation powered by SAS Viya, you have access to advanced profiling and data discovery features. These allow you to investigate and discover data quality issues you might want to address before using the data set in a report or analytical model.

In SAS Data Explorer, you can manage your data by viewing details about it. On the *Available* tab, you can search for the in-memory data set that you want to work with. On the *Details* tab shown in Figure 2, you can view some basic information about that table such as its column names and data types, when it was last profiled, the total number of columns, row count, and table size.


#	Name	Label	Type	Raw Len...	Formatted Length	Format	Tags
1	ID		double	8	12		
2	NAME		varchar	15	15		
3	COUNTRY		varchar	13	13		
4	EMAIL_ADDRESS		varchar	23	23		

**Figure 2. SAS Data Explorer: Details Tab**

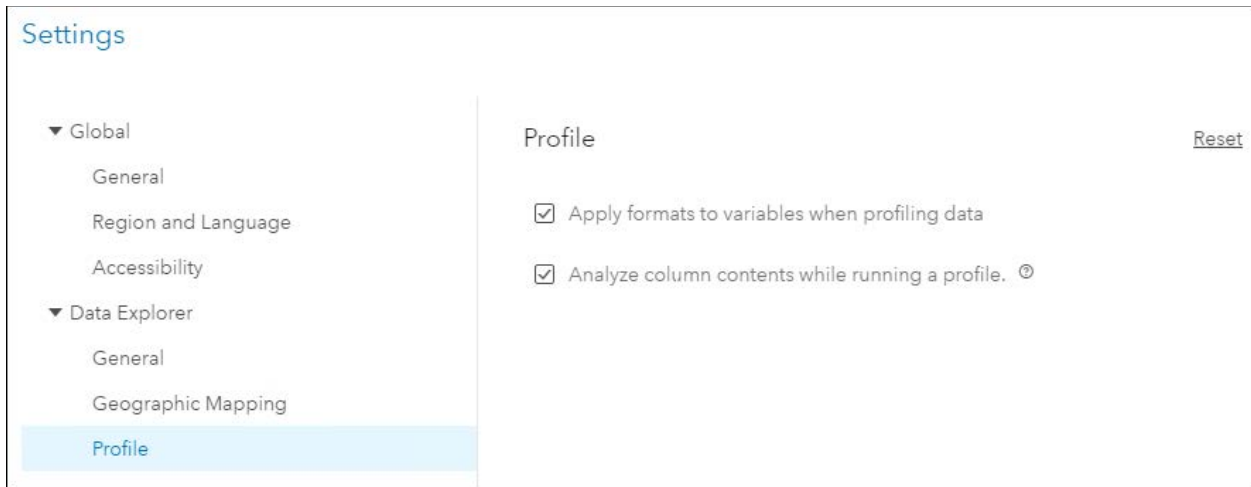
The *Sample Data* tab shown in Figure 3 helps you determine if this is the table you want to work with. It displays a sampling of data from the selected table. By default, the sample size is 100 rows; however, you can change this number and select the refresh button to view a different sample size.

ID	NAME	COUNTRY	EMAIL_ADDRESS
103	Michael Jameson	United States	michael.jameson@xxx.com
104	Albert Moore	US	AL.Moore@xxx.com
105	Harvey Monk	United States	harvey.monk@xxx.com
106	Susan Holmes	USA	sholmes@xxx.com
101	CARLOS VIERA	US	CARLOS.VIERA@XXX.COM
102	Richard Schmidt	US	richard.schmidt@xxx.com
107	Al Moore	US	al.moore@xxx.com
108	MIKE JAMESON	USA	Michael.Jameson@xxx.com
109	Sue Holmes	USA	sholmes@xxx.com
110	Anne Horton	US	Anne.Horton@XXX.com

**Figure 3. SAS Data Explorer: Sample Data Tab**

After reviewing the *Details* and *Sample Data* tabs, you can profile data to gain more insights on this data set. First, you should check your profile execution settings, by selecting  **Settings** next to your logon name on the browser to open the *Settings* window shown in Figure 4 below and then select **Profile** in the *Data Explorer* section. You can select the following profile options:

- Apply formats to variables when profiling data.
- Analyze column contents when running a profile.



**Figure 4. SAS Data Explorer: Profile Settings**

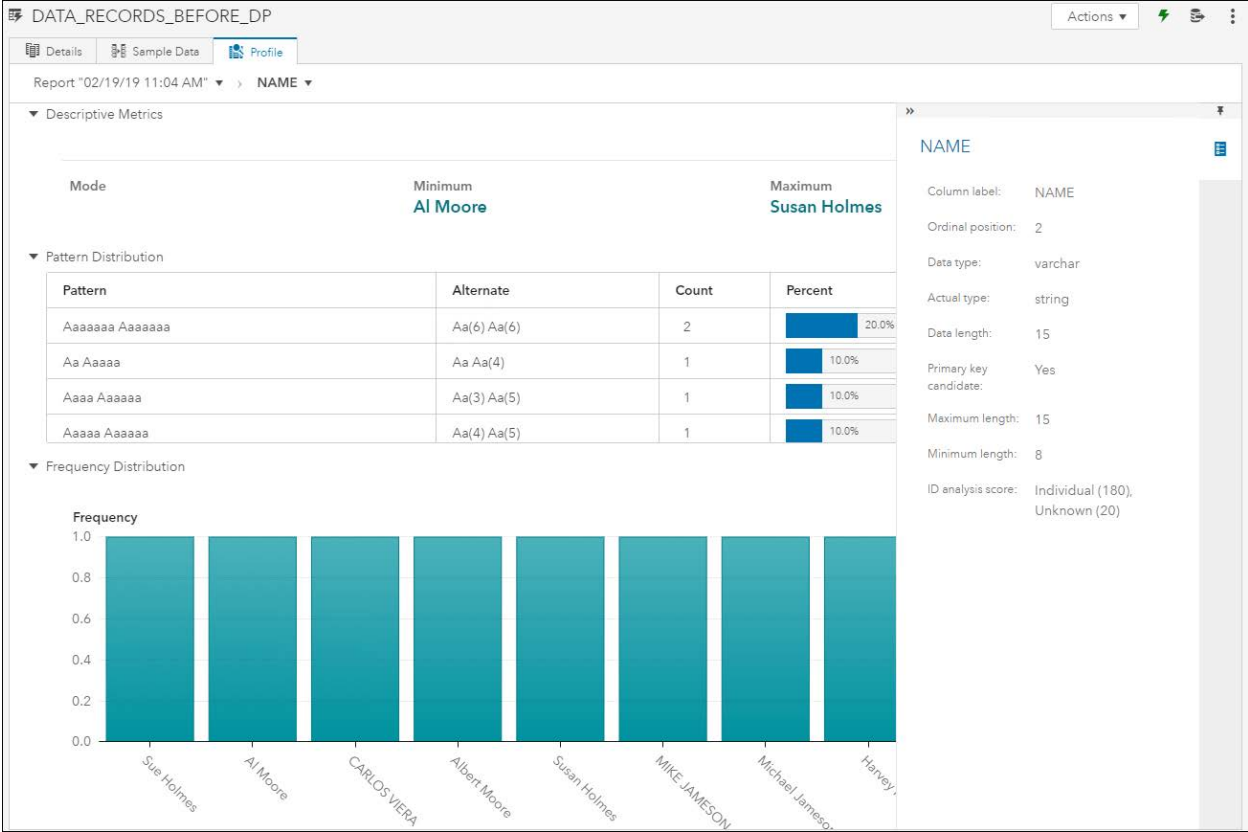
The first option, *Apply formats to variables when profiling data*, applies any SAS formats to the data when the profile is generated. The second option, *Analyze column contents when running a profile*, tags the column with a content type if one can be determined when profiling the data. For example, a column that contains street addresses might be tagged with the *Street Address* identifier. This analysis can impact the profiling performance.

Now that the profile settings are set, you can run the profile to get its metrics and content analysis information. If a data set has been profiled, it is displayed on the *Profile* tab. This data set has not been profiled yet; therefore, to run the profile, select the **Run Profile** button on the *Profile* tab, as shown in Figure 5. Once the profile has completed, the results are displayed on the *Profile* tab. It might take several seconds (or maybe even a minute or two) to calculate all the profile results depending on the size of the data set.

Column	Unique	Null	Blank	Pattern Count	Mean	Median	Mode	Standard Devi...	Standard Er...
COUNTRY	30.00% ...			3			US		
EMAIL_ADDRESS	90.00% (9)			8			sholm...		
ID	100.00% (10)				105.50	105.50		3.03	0.96
NAME	100.00% (10)			9					

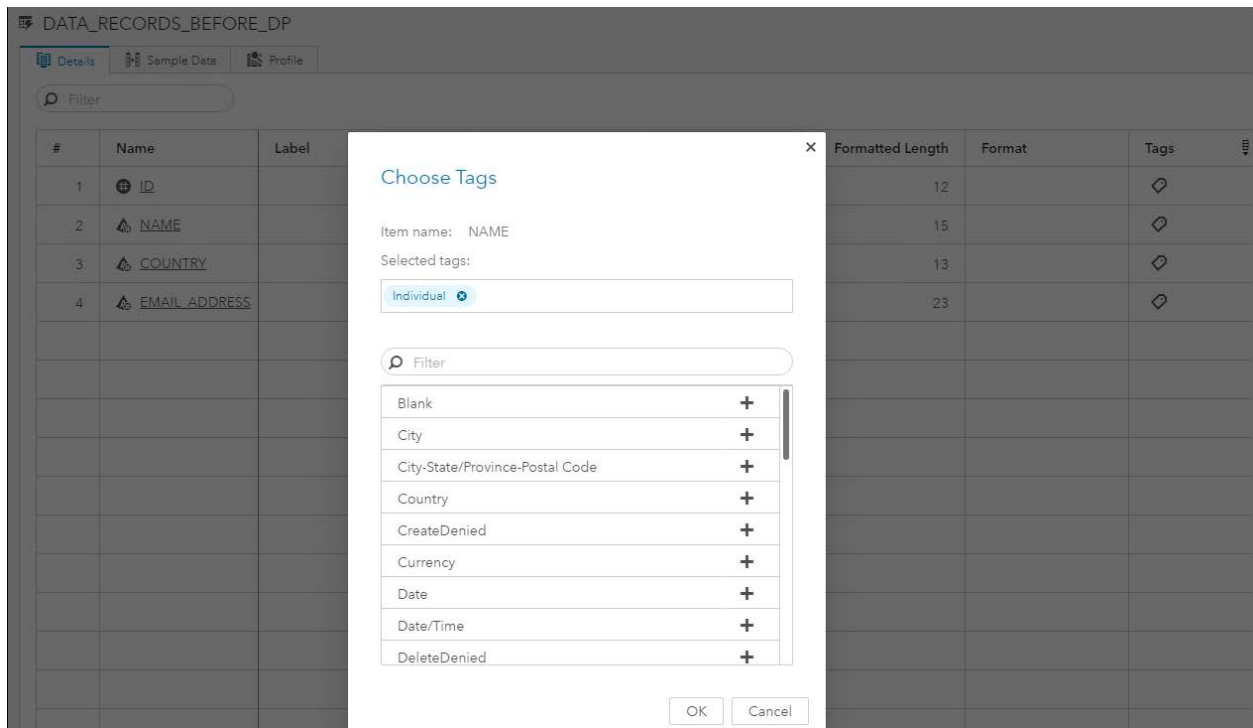
**Figure 5. SAS Data Explorer: Profile Tab**

A data profile report enables you to recognize data patterns, identify scarcity in the data, and review basic statistics for the selected table such as Mean, Median, and Mode and counts such as Null, Blank, and Pattern Counts. Some metrics might not be applicable to a column's data type. You can also drill-down into a particular column to view its specific metrics, frequency distributions, and column ID analysis results. Figure 6 represents the column profile results for the **Name** field.



**Figure 6. Name Column Profile Results**

For a pattern, an uppercase **A** represents an uppercase letter, a lowercase **a** represents a lowercase letter, a **9** represents a digit, and punctuation and spacing are displayed as-is. The highest ID analysis score for the column is the value with which the column is tagged. In Figure 7, the ID analysis identifies the *Name* column to contain *Individual* name data. This can be useful for identifying personal data that might need to be controlled or hidden. The column tag(s) are visible on the SAS Data Explorer – *Details* tab by selecting the tag button in the *Tags* column. You can also add additional column tags here.



**Figure 7. Name Column Tags**

Based on reviewing this data set, you might want to prepare the data as follows prior to using it in any reports or analytical models:

- Remove duplicate records from the data set.
- Standardize the *Country* column.
- Use consistent casing format for the *Name* and *Email\_Address* columns.

You also might want to augment this data set by adding a *Gender* field and parsing the *Name* field into *First Name* and *Last Name* columns. This can be accomplished by creating a SAS Data Studio plan file to transform the data set. Figure 8 depicts the desired data set after all the data preparation transformations in SAS Data Studio.

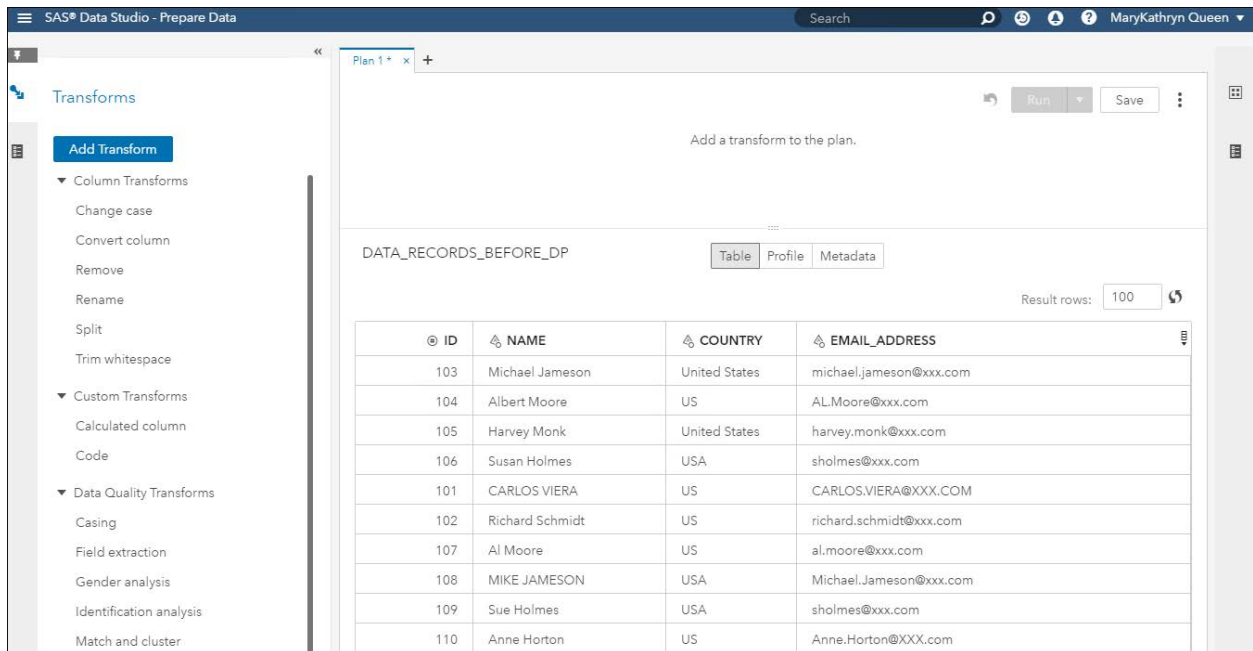
The screenshot shows the 'DATA\_RECORDS\_AFTER\_DP' dataset in SAS Data Studio. The table displays columns: ID, FIRST\_NAME, LAST\_NAME, GENDER, EMAIL\_ADDRESS, and COUNTRY. The 'Sample rows' are set to 100.

ID	FIRST_NAME	LAST_NAME	GENDER	EMAIL_ADDRESS	COUNTRY
105	Harvey	Monk	M	harvey.monk@xxx.com	USA
110	Anne	Horton	F	anne.horton@xxx.com	USA
106	Susan	Holmes	F	sholmes@xxx.com	USA
101	Carlos	Viera	M	carlos.viera@xxx.com	USA
103	Michael	Jameson	M	michael.jameson@xxx.com	USA
104	Albert	Moore	M	al.moore@xxx.com	USA
102	Richard	Schmidt	M	richard.schmidt@xxx.com	USA

**Figure 8. Data Records After Data Preparation**

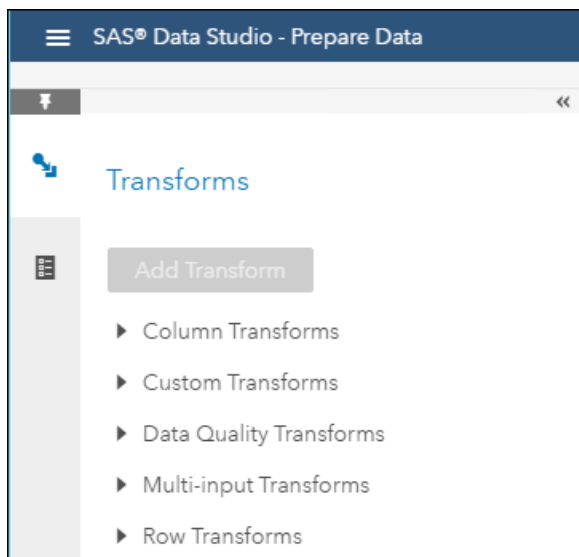
## DATA STUDIO PLANS

To create a SAS Data Studio plan based on the data set you are analyzing, you can select **Actions** → **Prepare data** in SAS Data Explorer while viewing the available in-memory table for the data set. As shown in Figure 9, this opens SAS Data Studio and starts a data plan based on the selected data set.



**Figure 9. SAS Data Studio Plan**

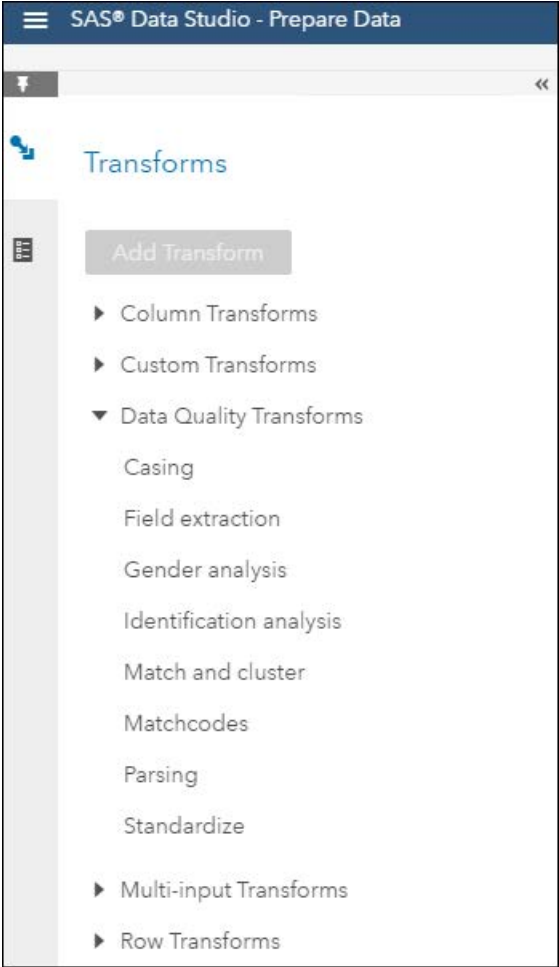
There are five categories of transforms that you can apply to your data set: Column, Custom, Data Quality, Multi-input, and Row. These categories are shown in Figure 10.



**Figure 10. SAS Data Studio: Transform Categories**

The Data Quality Transforms, shown in Figure 11, are powered by the SAS® Quality Knowledge Base (QKB). The QKB is a set of files containing proprietary rules, expressions,

and reference data that are combined to analyze and transform text data in various SAS products such as SAS® Data Integration Studio, DataFlux Data Management Studio, SAS® Event Stream Processing, and SAS Data Preparation. The data quality transforms apply a QKB locale (language and country combination) and a definition to a selected source column. Definitions define data formats for specific types of content and data cleansing. For example, a parse definition for a name describes how a name can be parsed into identifiable segments such as given name, middle name, and family name.



**Figure 11. SAS Data Studio: Data Quality Transforms**

To begin transforming the data set, first you might want to remove duplicate records. To facilitate this, generate **Matchcodes** for the columns you want to fuzzy match. In Figure 12, Matchcodes are generated for the *Name* and *Email\_Address* columns using the appropriate QKB definitions for the respective data types of the columns. The same Matchcode is generated for similar items so that they can be fuzzy matched. For example, when using the *Name* Matchcode definition, the same code is generated for the values "Michael Smith" and "Mike Smith".

Plan 1 \* x +

Matchcodes - Step 1 of 1 Run Save

1 Matchcodes

Source column: NAME Name of new column: NAME\_MATCH Locale: English (United States) Definition: Name Sensitivity: 85 Character length: 50

Source column: EMAIL\_ADD... Name of new column: EMAIL\_MATCH Locale: English (United Sta... Definition: E-mail Sensitivity: 85 Character length: 50

Options for new columns

---

DATA\_RECORDS\_BEFORE\_DP (session) Table Profile Metadata

Result rows: 100

NAME_MATCH	EMAIL_MATCH	ID	NAME	COUNTRY	EMAIL_ADDRESS
C&B&4&B\$\$\$\$\$\$B73...	B7J2&_WC&B_4@P\$\$\$\$...	103	Michael Jameson	United States	michael.jameson@xxx.c...
B&Y\$\$\$\$\$\$\$\$\$\$W...	&WB@Y_\$\$\$\$\$\$\$\$\$...	104	Albert Moore	US	AL.Moore@xxx.com
B&B3\$\$\$\$\$\$\$\$\$2YV\$...	2&YV_RB@P3\$\$\$\$\$\$\$\$\$...	105	Harvey Monk	United States	harvey.monk@xxx.com
2&WB&4\$\$\$\$\$\$\$\$\$4#_...	42@WB_4\$\$\$\$\$\$\$\$\$...	106	Susan Holmes	USA	sholmes@xxx.com
V&&Y&\$\$\$\$\$\$\$\$\$3&Y...	J&YW@4V7_Y&\$\$\$\$\$\$\$\$\$...	101	CARLOS VIERA	US	CARLOS.VIERA@XXX.C...
42B&~\$\$\$\$\$\$\$\$\$Y7J2...	Y7J2&Y84J2B78~\$\$\$\$\$...	102	Richard Schmidt	US	richard.schmidt@xxx.com
B&Y\$\$\$\$\$\$\$\$\$\$W...	&WB@Y_\$\$\$\$\$\$\$\$\$...	107	Al Moore	US	al.moore@xxx.com
C&B&4&B\$\$\$\$\$\$\$\$\$B73...	B7J2&_WC&B_4@P\$\$\$\$...	108	MIKE JAMESON	USA	Michael.Jameson@xxx.c...
2&WB&4\$\$\$\$\$\$\$\$\$4#_...	42@WB_4\$\$\$\$\$\$\$\$\$...	109	Sue Holmes	USA	sholmes@xxx.com
2&Y~&B\$\$\$\$\$\$\$\$\$&PP...	&P_2@Y~@P\$\$\$\$\$\$\$\$\$...	110	Anne Horton	US	Anne.Horton@XXX.com

**Figure 12. SAS Data Studio Plan: Matchcodes**

Once you have Matchcodes for the columns you want to fuzzy match, the next step is to **Standardize** columns that you want to match exactly. In Figure 13, the *Country* column is standardized using the appropriate QKB definition for its data type.



Standardize - Step 2 of 2

1 Matchcodes — 2 Standardize

Source column: COUNTRY Name of new column: COUNTRY\_STND Locale: English (United States) Definition: Country (ISO 3 Char) Character length: 3

Options for new columns

DATA\_RECORDS\_BEFORE\_DP (session) Table Profile Metadata

Result rows: 100

COUNTRY_STND	NAME_MATCH	EMAIL_MATCH	ID	NAME	COUNTRY	EMAIL_ADDRESS
USA	C&B&4&B\$\$\$\$\$...	B7J2&_WC&B_4@...	103	Michael Jameson	United States	michael.jameson...
USA	B&Y\$\$\$\$\$\$\$\$\$...	&WB@Y_\$\$\$\$\$\$\$...	104	Albert Moore	US	AL.Moore@xxx.com
USA	B&B3\$\$\$\$\$\$\$\$\$...	2&YV_RB@P3\$\$\$\$\$...	105	Harvey Monk	United States	harvey.monk@xxx....
USA	2&WB&4\$\$\$\$\$\$\$\$\$...	42@WB_4\$\$\$\$\$\$\$...	106	Susan Holmes	USA	sholmes@xxx.com
USA	V&&Y\$\$\$\$\$\$\$\$\$...	J&YW@4V7_Y&\$\$...	101	CARLOS VIERA	US	CARLOS.VIERA@X...
USA	42B&~\$\$\$\$\$\$\$\$\$...	Y7J2&Y84J2B78~...	102	Richard Schmidt	US	richard.schmidt@x...
USA	B&Y\$\$\$\$\$\$\$\$\$...	&WB@Y_\$\$\$\$\$\$\$...	107	Al Moore	US	al.moore@xxx.com
USA	C&B&4&B\$\$\$\$\$...	B7J2&_WC&B_4@...	108	MIKE JAMESON	USA	Michael.Jameson...
USA	2&WB&4\$\$\$\$\$\$\$\$\$...	42@WB_4\$\$\$\$\$\$\$...	109	Sue Holmes	USA	sholmes@xxx.com
USA	2&Y~&B\$\$\$\$\$\$\$\$\$...	&P_2@Y~@P\$\$\$\$\$...	110	Anne Horton	US	Anne.Horton@XX...

**Figure 13. SAS Data Studio Plan: Standardize**

Now that you have the components to aid in finding duplicate records in your data set, you can **Match and Cluster** the records. In Figure 12, you want to Match and Cluster records based on the rule where the *Name\_Match*, *Email\_Match*, and *Country\_Stnd* columns are the same. When this is the case, the records receive the same *Cluster\_ID* value.

Plan 1\* x +

Match and Cluster - Step 3 of 3

1 Matchcodes — 2 Standardize — 3 Match and Cluster

New column name: Cluster\_ID

Any of the following:

▼ Rule 1 of 1

Column: NAME\_MATCH

AND EMAIL\_MATCH

AND COUNTRY\_STND

Interpret empty strings as null values

Allow null values to match

Exclude rows from matching if the value of the following column is "True".

Column: Select an item

DATA\_RECORDS\_BEFORE\_DP (session)

Table Profile Metadata

Result rows: 100

COUNTRY_STND	NAME_MATCH	EMAIL_MATCH	ID	NAME	COUNTRY	EMAIL_ADD	Cluster_ID
USA	C&B&4&B\$\$\$\$\$\$B...	\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$	103	Michael Jam...	United States	michael.ja...	AwAAAAAAAAAB...
USA	B&Y\$\$\$\$\$\$\$\$\$&...	\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$	104	Albert Moore	US	AL.Moore...	AwAAAAAAAAA...
USA	B&B3\$\$\$\$\$\$\$\$\$2...	\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$	105	Harvey Monk	United States	harvey.m...	AQAAAAAAAAA...

**Figure 14. SAS Data Studio Plan: Match and Cluster**

With the data records clustered, you can remove the duplicate records by using the **Code** transform in the Custom Transforms section. Figure 15 shows the code that will keep the record with the lowest ID value for the cluster.

Plan 1\* x +

Code - Step 4 of 4

1 Matchcodes — 2 Standardize — 3 Match and Cluster — 4 Code

DATA step

```

1 data {{_dp_outputTable}} (caslib={{_dp_outputCaslib}});
2 set {{_dp_inputTable}} (caslib={{_dp_inputCaslib}});
3 by Cluster_ID ID; /* group by Cluster_ID, then by ID value */
4 if first.Cluster_ID then output; /* returns first grouping */
5 run;

```

DATA\_RECORDS\_BEFORE\_DP (session)

Table Profile Metadata

Result rows: 100

COUNTRY_STND	NAME_MATCH	EMAIL_MATCH	ID	NAME	COUNTRY	EMAIL_ADD	Cluster_ID
USA	B&B3\$\$\$\$\$...	\$\$\$\$\$\$\$\$\$...	105	Harvey Monk	United States	harvey.mo...	AQAAAAAAAAA...
USA	2&Y~&B\$\$\$\$\$...	\$\$\$\$\$\$\$\$\$...	110	Anne Horton	US	Anne.Hort...	AwAAAAAAAAAD...
USA	42B&~\$\$\$\$\$...	\$\$\$\$\$\$\$\$\$...	102	Richard Schm...	US	richard.sc...	AgAAAAAAAAAB...
USA	2&WB&4\$\$\$\$\$...	\$\$\$\$\$\$\$\$\$...	106	Susan Holmes	USA	sholmes@...	AQAAAAAAAAAB...
USA	V&Y&\$\$\$\$\$...	\$\$\$\$\$\$\$\$\$...	101	CARLOS VIERA	US	CARLOS.V...	AgAAAAAAAAAA...
USA	B&Y\$\$\$\$\$...	\$\$\$\$\$\$\$\$\$...	104	Albert Moore	US	AL.Moore...	AwAAAAAAAAAA...
USA	C&B&4&B\$\$\$\$\$...	\$\$\$\$\$\$\$\$\$...	103	Michael Jam...	United States	michael.ja...	AwAAAAAAAAAB...

**Figure 15. SAS Data Studio Plan: Code for Removing Duplicate Records**

Now that the duplicate records have been removed, you can perform a **Gender Analysis** on the *Name* column using a QKB definition to make a best guess of the gender for the supplied name, as shown in Figure 16.

Plan 1\* x +

Gender Analysis - Step 5 of 5

1 Matchcodes — 2 Standardize — 3 Match and Cluster — 4 Code — 5 Gender Analysis

Source column: NAME Name of new column: GENDER Locale: English (United States) Definition: Name Character length: 1

Options for new columns

DATA\_RECORDS\_BEFORE\_DP (session)

Table Profile Metadata

Result rows: 100

GENDER	COUNTRY_STI	NAME_MATC	EMAIL_MATC	ID	NAME	COUNTRY	EMAIL_ADDR	Cluster_ID
M	USA	B&B3\$\$\$\$\$...	2&YV_RB@P...	105	Harvey Monk	United States	harvey.mon...	AQAAAAA...
F	USA	2&Y~&B\$\$\$\$\$...	&P_2@Y~@...	110	Anne Horton	US	Anne.Horto...	AwAAAAA...
M	USA	C&B&4&B\$...	B7J2&_WC...	103	Michael Jam...	United States	michael.jam...	AAAAAAA...
M	USA	B&Y\$\$\$\$\$...	&WB@Y_\$\$...	104	Albert Moore	US	AL.Moore@...	AAAAAAA...
M	USA	42B&~\$\$\$\$\$...	Y7J2&Y84J...	102	Richard Sch...	US	richard.sch...	AgAAAAA...
F	USA	2&WB&4\$5\$...	42@WB_4\$...	106	Susan Holmes	USA	sholmes@xx...	AQAAAAA...
M	USA	V&Y&\$\$\$\$\$...	J&YW@4V7...	101	CARLOS VIE...	US	CARLOS.VIE...	AgAAAAA...

**Figure 16. SAS Data Studio Plan: Gender Analysis**

Next, you can use the **Casing** transform to propercase the *Name* column and lowercase the *Email\_Address* column in order to put them in a consistent format, as shown in Figure 17.

Plan 1\* x +

Casing - Step 6 of 6 Run Save

1 Matchcodes — 2 Standardize — 3 Match and Cluster — 4 Code — 5 Gender Analysis — 6 Casing

Source column: NAME Name of new column: NAME\_CASE Locale: English (United States) Casing: Proper (Name) Character length: 50

Source column: EMAIL\_ADD... Name of new column: EMAIL\_ADDRESS\_CA... Locale: English (United States) Casing: Lower Character length: 50

Options for new columns

DATA\_RECORDS\_BEFORE\_DP (session) Table Profile Metadata Result rows: 100

NAME_CASE	EMAIL_ADDRESS_CASE	GENDER	COUNTRY	NAME_M	EMAIL_M	ID	NAME	COUNT	EMAIL_A
Harvey Monk	harvey.monk@xxx.com	M	USA	B&B3\$...	2&VY_R...	105	Harvey ...	Unite...	harvey...
Anne Horton	anne.horton@xxx.com	F	USA	2&Y~&B...	&P_2@Y...	110	Anne Ho...	US	Anne.Ho...
Michael James...	michael.jameson@xxx.com	M	USA	C&B&4&...	B7J2&_...	103	Michael ...	Unite...	michael...
Albert Moore	al.moore@xxx.com	M	USA	B&Y\$S\$...	&WB@Y...	104	Albert M...	US	AL.Moor...
Richard Schmidt	richard.schmidt@xxx.com	M	USA	42B&~\$...	Y7J2&Y...	102	Richard ...	US	richard.s...
Susan Holmes	sholmes@xxx.com	F	USA	2&WB&...	42@WB_...	106	Susan H...	USA	sholmes...
Carlos Viera	carlos.viera@xxx.com	M	USA	V&Y&\$...	J&YW@...	101	CARLOS...	US	CARLOS...

**Figure 17. SAS Data Studio Plan: Casing**

To split the *Name\_Case* column into *First\_Name* and *Last\_Name* columns, you can use the **Parsing** transform with the appropriate QKB definition for the column's data type, as shown in Figure 18.

Plan 1\* x +

Parsing - Step 7 of 7

Run Save

1 Standardize 2 Standardize 3 Match and Cluster 4 Code 5 Gender Analysis 6 Casing 7 Parsing

Source column: NAME\_CASE Locale: English (United States) Definition: Name

Selected tokens: Family Name Given Name

Available tokens (4):

- Middle Name  
NAME\_CASE\_MiddleName\_Parse
- Prefix  
NAME\_CASE\_Prefix\_Parse

Selected tokens (2):

- Given Name  
FIRST\_NAME
- Family Name  
LAST\_NAME

Options for new columns

DATA\_RECORDS\_BEFORE\_DP (session)

Table Profile Metadata

Result rows: 100

FIRST_NAME	LAST_NAME	NAME_CASE	EMAIL_A	GENDER	COUNTRY	NAME_M	EMAIL_M	ID	NAME	CO
Harvey	Monk	Harvey Monk	harvey...	M	USA	B&B3\$...	2&YV_R...	105	Harvey ...	United ..
Anne	Horton	Anne Horton	anne.ho...	F	USA	2&Y~&...	&P_2@Y...	110	Anne H...	US
Michael	Jameson	Michael Ja...	michael...	M	USA	C&B&4...	B7J2&_...	103	Michael...	United ..
Albert	Moore	Albert Moore	al.moor...	M	USA	B&Y\$3\$...	&WB@Y...	104	Albert ...	US
Richard	Schmidt	Richard Sch...	richard...	M	USA	42B&~\$...	Y7J2&Y...	102	Richard ...	US
Susan	Holmes	Susan Holmes	sholme...	F	USA	2&WB&...	42@WB...	106	Susan H...	USA
Carlos	Viera	Carlos Viera	carlos.vi...	M	USA	V&&Y&...	J&YW@...	101	CARLO...	US

**Figure 18. SAS Data Studio Plan: Parsing**

Finally, you can use the **Code** transform to rearrange and only return the columns desired for your prepared data table that can then be used in your reports or analytical models. Figure 19 contains the code to create the desired output table for the data set.

Plan 1\* x +

Code - Step 8 of 8 Run Save

Standardize — 3 Match and Cluster — 4 Code — 5 Gender Analysis — 6 Casing — 7 Parsing — 8 Code

CASL

```

1 /* Rearrange column order for output table */
2 queryCode='create table ""||_dp_outputCaslib ||"".""||_dp_outputTable||"" {options replace=true}
3 as select ID, FIRST_NAME, LAST_NAME, GENDER, EMAIL_ADDRESS_CASE as "EMAIL_ADDRESS", COUNTRY_STND as "COUNTRY"
4 from ""||_dp_inputCaslib||"".""||_dp_inputTable ||""';
5 print queryCode;
6 fedSQL.execDirect / query=queryCode;

```

DATA\_RECORDS\_BEFORE\_DP (session) Table Profile Metadata

Result rows: 100

ID	FIRST_NAME	LAST_NAME	GENDER	EMAIL_ADDRESS	COUNTRY
105	Harvey	Monk	M	harvey.monk@xxx.com	USA
110	Anne	Horton	F	anne.horton@xxx.com	USA
103	Michael	Jameson	M	michael.jameson@xxx.com	USA
104	Albert	Moore	M	al.moore@xxx.com	USA
102	Richard	Schmidt	M	richard.schmidt@xxx.com	USA
106	Susan	Holmes	F	sholmes@xxx.com	USA
101	Carlos	Viera	M	carlos.viera@xxx.com	USA

**Figure 19. SAS Data Studio Plan: Code to Rearrange Column Order**

The plan file and the resulting table can be saved, as shown in Figure 20.

Save As

Search

SAS Content > My Folder

- My Favorites
- My Folder
- ESP Projects
- gelcontent
- Model Repositories
- Products
- Projects
- SAS Videos

Name: Prepare Data Records Type: Data plan

Save plan and table  Save plan  Save table

Table name: DATA\_RECORDS\_AFTER\_DP Label: Enter label Library: cas-shared-default/DM

If the name of the target table already exists:  Cancel save  Replace table

Save Cancel

**Figure 20. SAS Data Studio Plan: Save Plan and Table**

Figure 21 is the resulting saved table from the data plan used to prepare the before data set.

The screenshot shows the SAS Data Explorer interface. The main window displays a table titled 'DATA\_RECORDS\_AFTER\_DP'. The table has the following columns: ID, FIRST\_NAME, LAST\_NAME, GENDER, EMAIL\_ADDRESS, and COUNTRY. The data rows are as follows:

ID	FIRST_NAME	LAST_NAME	GENDER	EMAIL_ADDRESS	COUNTRY
105	Harvey	Monk	M	harvey.monk@xxx.com	USA
110	Anne	Horton	F	anne.horton@xxx.com	USA
106	Susan	Holmes	F	sholmes@xxx.com	USA
101	Carlos	Viera	M	carlos.viera@xxx.com	USA
103	Michael	Jameson	M	michael.jameson@xxx.com	USA
104	Albert	Moore	M	al.moore@xxx.com	USA
102	Richard	Schmidt	M	richard.schmidt@xxx.com	USA

**Figure 21. Saved Table After Data Preparation**

If needed, you also have the option to create a job from the plan file that can then be scheduled to run according to some time-based event. For example, you could schedule the plan job to execute and replace the resulting saved table every day at midnight.

## CONCLUSION

Using the advanced data profiling and data discovery techniques in SAS Data Explorer helps you determine the data cleansing needs for your data set. Then, with SAS Data Studio, you can implement your data cleansing plan using transforms, which include data quality transforms that use the SAS Quality Knowledge Base (QKB). SAS Data Preparation powered by SAS Viya helps ensure that your data is cleansed and ready for use in your reports and analytical models to better meet your business needs and aid in your business decisions.

## RESOURCES

- SAS® Viya® 3.4: Data Preparation / Getting Started. Available at <https://go.documentation.sas.com/?cdcid=dprepcdc&cdcVersion=2.2&docsetId=dprepgs&docsetTarget=home.htm&locale=en>
- Rausch, Nancy. 2018. "What's New in SAS Data Management." *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC: SAS Institute Inc. Available at <https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1669-2018.pdf>
- Rineer, Brian 2018. "Doin' Data Quality in SAS® Viya®" *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. Available at <https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2156-2018.pdf>
- SAS® Data Management Community. Available at [https://communities.sas.com/t5/Data-Management/ct-p/data\\_management](https://communities.sas.com/t5/Data-Management/ct-p/data_management)

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mary Kathryn Queen

SAS Institute Inc.

[MaryKathryn.Queen@sas.com](mailto:MaryKathryn.Queen@sas.com)

<http://www.sas.com>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.