SAS2156-2018

Doin' Data Quality in SAS® Viya®

Brian Rineer, SAS Institute Inc

ABSTRACT

SAS® Viya® introduces data quality capabilities for big data through Data Preparation and DATA step programming for SAS® Cloud Analytic Services (CAS). In this session, a Senior Software Development Manager at SAS shows how to configure SAS® Data Quality transformations in SAS® Data Studio and how to submit DATA step functions that are created in SAS Data Quality for execution in CAS. We also cover management of the vital SAS® Quality Knowledge Base in SAS® Environment Manager.

INTRODUCTION

Data quality capabilities long available in SAS® Data Quality Server are now implemented in SAS Viya. SAS Data Quality in SAS Viya provides support for big data quality with the distributed processing power of CAS.

SAS Data Quality enables analysis and cleansing of structured text data. SAS Data Quality operations help you discover the semantic types of your data, break down data into constituent parts, standardize records into consistent formats, and identify potential duplicates.

If you're a SAS Data Quality Server user, you'll recognize a familiar set of data quality operations in the SAS Data Quality offering in SAS Viya. As with SAS Data Quality Server, you can analyze and categorize your data and cleanse it in preparation for analytics and investigation, or simply for better data hygiene and cleaner reporting. If you're new to data quality, see *SAS® Data Quality: Getting Started* for an overview and examples of SAS Data Quality operations.

In this paper, we briefly show how these data quality operations are accessed in SAS Viya.

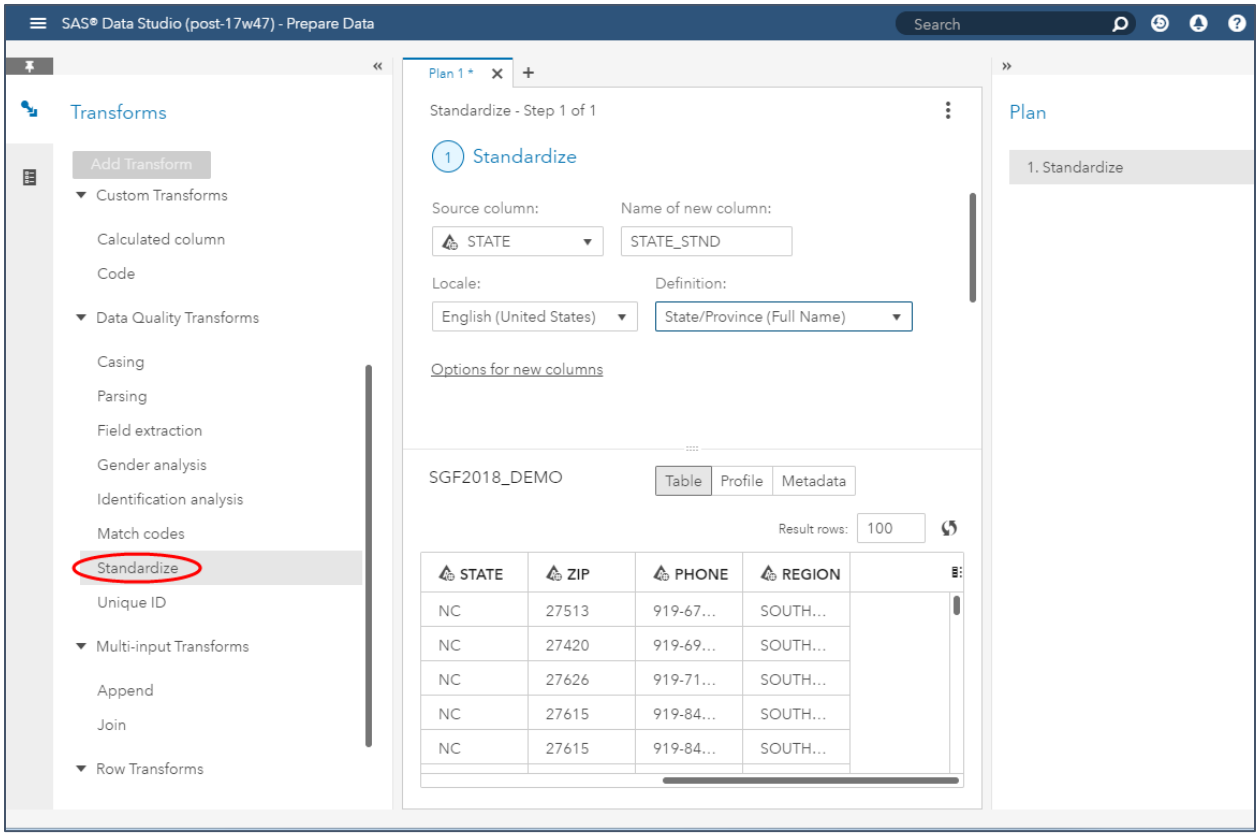
DATA QUALITY IN SAS DATA STUDIO

SAS Data Studio is a SAS® Data Preparation application in which you define transforms that are applied to your data in CAS. You can specify transforms to prepare your data for your analytics pipeline or to meet whatever data cleansing you might have.

In SAS Data Studio, you'll define transforms with an easy-to-use web browser interface. This interface enables you to create and modify transforms any time, in any environment. Once created, the set of transforms you've defined constitute a "plan" that is submitted for execution in CAS.

The data quality operations that are available in SAS Data Studio mirror the data quality operations available in traditional SAS products such as SAS Data Quality Server and DataFlux® Data Management Studio. These operations include identification analysis, standardization, parsing, extraction, casing, gender analysis, and matchcode generation.

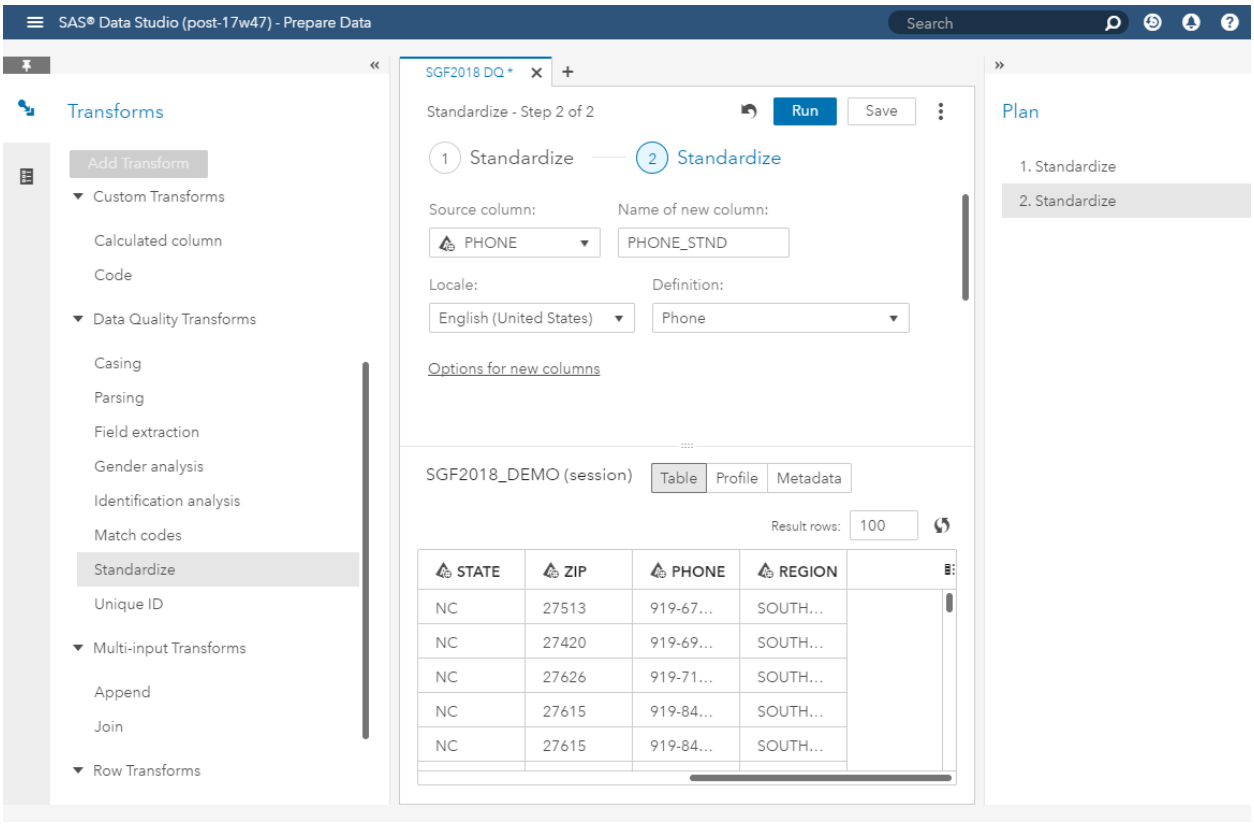
An example of a data quality transform created in SAS Data Studio is shown below. In the example, we create a simple standardization transform. We do this by choosing to add a **Standardize** transform from the side menu, clicking to select a column to transform, and then specifying a context known as a "definition" and a locale for that definition.



Display 1: A Standardize Transform Applied to a STATE Column in SAS Data Studio

In our example, we specify that we will standardize the values in a STATE column. We select the **State/Province (Full Name)** definition and the **English, United Stat**es locale to inform the software that the values to be transformed are names of US states.

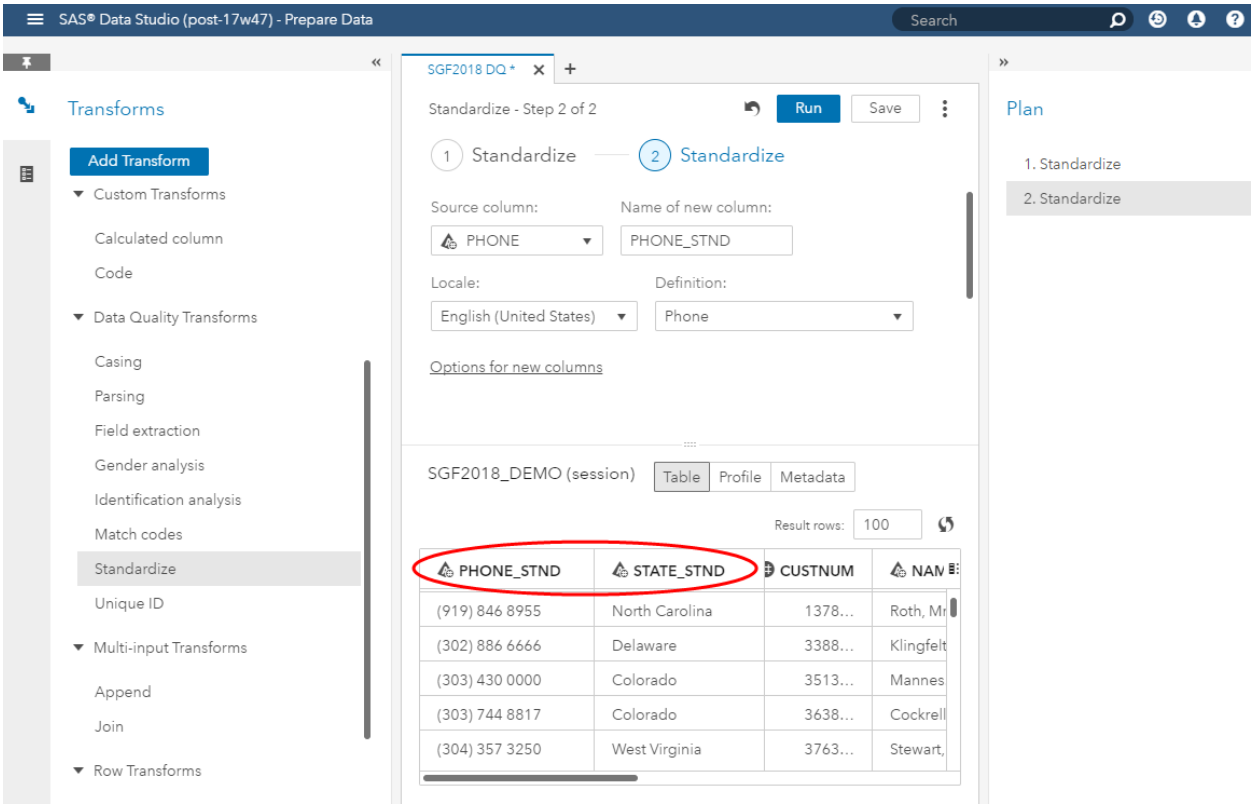
To standardize another column, we add a second transform to the plan:



Display 2: A Second Standardize Transform Added to a Plan in SAS Data Studio

In this example, we add a transform to standardize the PHONE column. Notice that the plan now contains two Standardize transforms.

The plan can contain many transforms that will be executed in series. When the complete plan is ready, we click **Run** to submit it to CAS for execution on the CAS massive parallel processing grid. When the plan has been run, we see a preview of the two new columns that contain the output of our transforms:



Display 3: Columns Created by Standardize Transforms in SAS Data Studio

For more information about SAS Data Studio and about SAS Data Studio transforms and plans, refer to *SAS Data Studio 2.1: User's Guide*.

DATA STEP

If you're a SAS Data Quality Server user, you'll be glad to know that the same data quality DATA step functions exist in the SAS Data Quality offering in SAS Viya as in SAS Data Quality Server, with the same familiar syntax. But with SAS Data Quality in SAS Viya, you can submit your DATA step program for execution in CAS.

To submit your DATA step code for execution in CAS, you'll use a CAS action called *dataStep.runCode*. The syntax for your functions will be the same as in SAS Data Quality Server. Here is an example:

```
data CUSTOMERS;
    set SGF2018_DEMO;
    STATE_STND = dqStandardize(STATE, 'State/Province (Full Name)');
run;
```

This code applies the **State/Province (Full Name)** definition to standardize data in the STATE variable in the SGF2018_DEMO data set.

To submit this step for execution in CAS, you can wrap it in a call to the *dataStep.runCode* CAS action, and submit the call by invoking the *cas* procedure:

```
proc cas;

action dataStep.runCode /
code="
data CUSTOMERS;
    set SGF2018_DEMO;
    STATE_STND = dqStandardize(STATE, 'State/Province (Full Name)');
     ";
run;
```

You can submit this program in SAS Studio or anywhere you can execute SAS code. When you submit the program in your SAS environment, your DATA step code is routed to CAS and your data is processed in CAS.

For information about the *dataStep.runCode* CAS action, see the documentation for the DATA Step CAS Action Set.

For information about SAS Data Quality DATA step functions, with examples and syntax, see *SAS® Data Quality 3.3* and *SAS® 9.4 Data Quality Server: Language Reference*. Be aware that at the time of this writing, not all data quality functions are available for execution in CAS. In particular, functions that apply schemes are not yet supported in CAS.

PERFORMANCE

Regardless of whether you create data quality transforms in SAS Data Studio or invoke data quality DATA step functions using the *datasStep.runCode* CAS action, you will be harnessing the power of the CAS massively parallel processing architecture.

In CAS, your data is distributed across multiple worker nodes on the CAS grid. Each worker node processes the records that are located on that node. The processing occurs in parallel, meaning that the time required to process a large data set with SAS Data Quality in SAS Viya is a fraction of the time required to process the same data set with SAS Data Quality Server. Also, performance scales linearly in CAS according to the number of worker nodes available in the system.

While the time needed to execute data quality operations in SAS Viya depends on many factors — size of data set, complexity of data, grid load, and so on — in internal tests, we have seen data quality operations executed on millions of records in a few seconds on a large CAS grid.

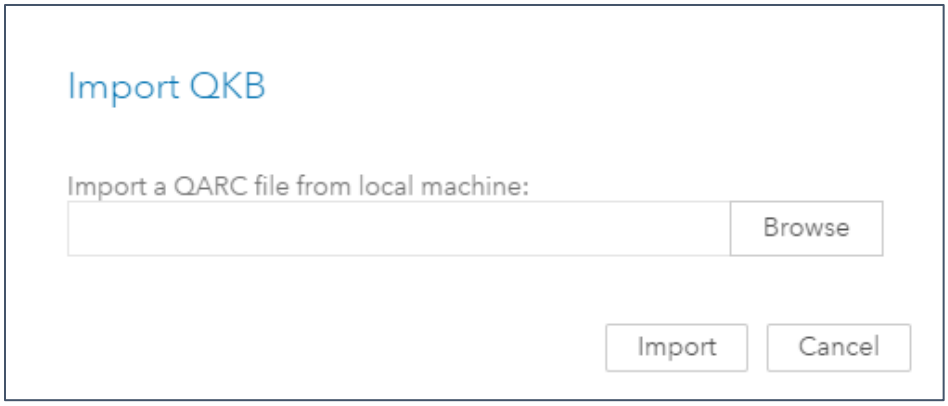
THE SAS QUALITY KNOWLEDGE BASE

As with SAS Data Quality Server, SAS Data Quality in SAS Viya requires a SAS Quality Knowledge Base (QKB). The QKB supplies rules and reference data used to analyze and transform your data.

The latest production version of the SAS® Quality Knowledge Base for Contact Information is deployed with SAS Data Quality. The entire QKB is deployed with your software order, with support for all available locales. At the time of this writing, this product supports thirty-nine locales. For more information about the SAS Quality Knowledge Base for Contact Information, refer to *About SAS Quality Knowledge Base* in the SAS Quality Knowledge Base for Contact Information online help.

To make the QKB available for use in CAS, you'll need to import it into your CAS system after you start the system for the first time. You can import a QKB into CAS using the SAS Environment Manager.

To import a QKB using SAS Environment Manager, open the SAS Home page in a web browser. Then, on the side menu, select **Reference Data** and then **Quality Knowledge Bases**. Next, on the Quality Knowledge Bases page, select **Import QKB**:



Display 4: Import QKB Dialog Box in SAS Environment Manager

The import process takes a few minutes. Generally, it takes a little longer for CAS systems with many worker nodes than for CAS systems with fewer worker nodes. After import, the QKB is stored as a repeated table in CAS. This means that a full copy of the QKB is available to every worker node for optimized processing at run time. For further instructions, refer to the QKB Management topic in *SAS Viya Administration*.

If you already have a QKB that you would like to begin using with SAS Data Quality in SAS Viya — for example, a customized QKB that you used previously with SAS Data Quality Server or the DataFlux® Data Management Platform — you can import that QKB into CAS as well. Note, however, that you must first convert your QKB into a format that is ready for import into CAS. This format is called a QKB Archive, or QARC. To create a QARC from your QKB, you'll use a utility that is provided with CAS. For details, see the QKB Management topic in *SAS Viya Administration*.

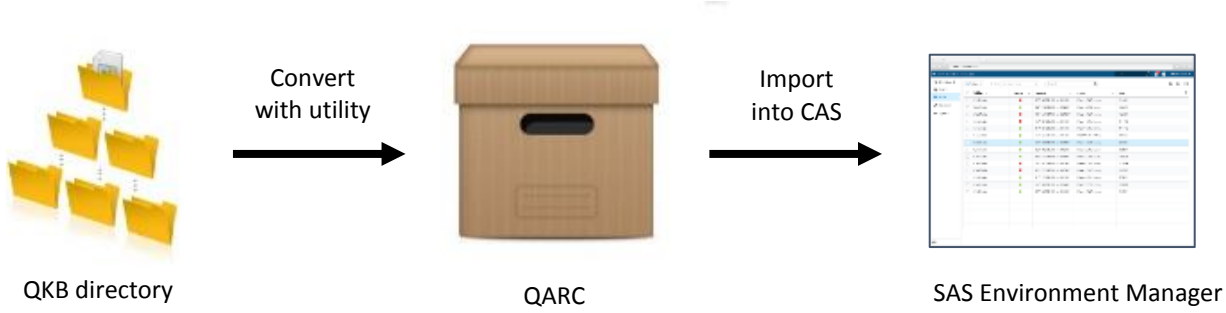
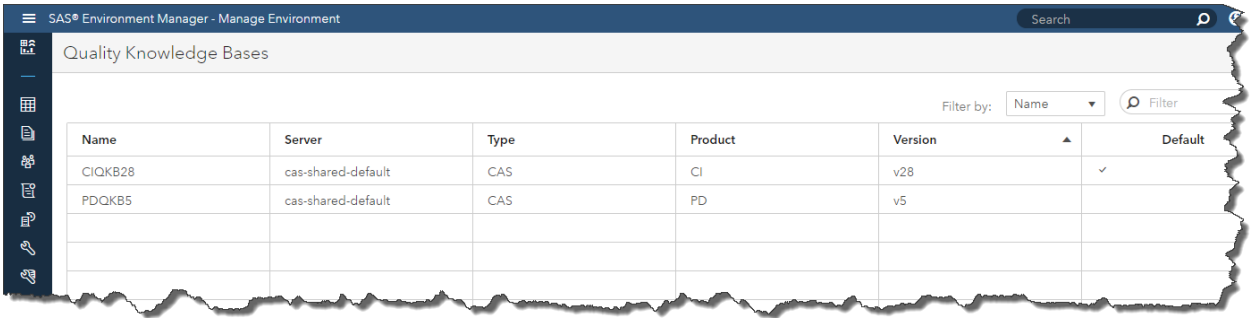


Figure 1: Process for Importing a Legacy QKB into CAS

After you have imported one or more QKBs into your CAS server, you can view the available QKBs using SAS Environment Manager:



Display 5: QKBs Viewed in SAS Environment Manager

Notice that there is a default QKB setting for your CAS server. If you open the properties screen for the default QKB, you can also find the default QKB locale setting. These settings tell SAS Data Studio which QKB you want to use when you execute data quality transforms, and which QKB and QKB locale you want to use when you submit data quality DATA step functions. (You can override these defaults in your DATA step program. See the *SAS Data Quality: Language Reference* for details.)

If you want to specify a default QKB and QKB locale, you can edit your CAS config file, and then restart your CAS server. Thereafter, you will see your default QKB and QKB locale settings in SAS Environment Manager. And the data quality transforms and DATA step functions will use these settings when they execute. For instructions on editing the default QKB and QKB locale settings in your CAS config file, see the QKB Management topic in the *SAS Viya Administration* documentation.

If you prefer to manage your QKBs programmatically rather than via a browser interface, you can call a CAS action to list QKBs, view QKB properties, and import or remove a QKB from CAS. An example is shown below. This example calls the *qkb.listQKBs* action to get a list of available QKBs for a given CAS server:

```
proc cas;
  action qkb.listQKBs;
  run;
quit;
```

For information about CAS actions for QKB management, refer to documentation for the QKB CAS action set.

GETTING STARTED

New to SAS Data Quality and eager to get started? Want a summary of SAS Data Quality functionality with links to references? If so, see *SAS® Data Quality: Getting Started*. This short book provides an overview of the various SAS Data Quality operations and what they do. It provides links to other books that contain product-specific details. It's a gateway to individual reference manuals that document the applications that surface SAS Data Quality operations and provide syntax and examples for programmatic interfaces.

CONCLUSION

With the introduction of SAS Data Quality in SAS Viya, you can harness the power of CAS to perform data quality operations on big data in a fraction of the time required by traditional data quality products. You can create data quality transforms with an easy-to-use web browser interface in SAS Data Studio, or write your own DATA step code to create custom programs. As always, a QKB is critical to SAS Data Quality. In SAS Viya, you can manage your QKBs via a web browser interface in SAS Environment Manager.

REFERENCES

DATA Step Action Set. Available at http://go.documentation.sas.com/?cdcId=vdmmlcdc&cdcVersion=8.11&docsetId=caspg&docsetTarget=cas-datastep-runcode.htm&locale=en.

QKB Action Set. Available at http://go.documentation.sas.com/?cdcId=dqcdc&cdcVersion=3.3&docsetId=casactdq&docsetTarget=cas-qkb-TblOfActions.htm&locale=en.

SAS® Data Quality 3.3: Getting Started. Available at http://go.documentation.sas.com/?cdcId=dqcdc&cdcVersion=3.3&docsetId=dqgs&docsetTarget=home.htm&locale=en.

SAS® Data Quality 3.3 and SAS® 9.4 Data Quality Server: Language Reference. Available at http://go.documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.3&docsetId=dqclref&docsetTarget=titlepage.htm&locale=en.

SAS® Quality Knowledge Base for Contact Info online Help. Available at http://support.sas.com/documentation/onlinedoc/qkb/28/QKBCI28/Help/qkb-help.html.

SAS® Data Studio 2.1: User's Guide. Available at http://go.documentation.sas.com/?cdcId=dprepcdc&cdcVersion=2.1&docsetId=datastudioadv&docsetTarget=titlepage.htm&locale=en.

SAS® Viya® 3.3 Administration: QKB Management. Available at http://go.documentation.sas.com/?cdcId=dqcdc&cdcVersion=3.3&docsetId=calqkb&docsetTarget=titlepage.htm&locale=en.

RECOMMENDED READING

- 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://support.sas.com/resources/papers/proceedings18/SAS1669-2018.pdf.
- Rineer, Brian. 2015. "Garbage In, Gourmet Out: How to Leverage the Power of the SAS® Quality Knowledge Base." *Proceedings of the SAS® Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings15/SAS1390-2015.pdf.
- Rineer, Brian. 2016. "Get out of DATA Step Code and into Quality Knowledge Bases." *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc. Available at https://support.sas.com/resources/papers/proceedings16/SAS5644-2016.pdf.
- *SAS Data Quality 3.3: Getting Started*. Available at http://go.documentation.sas.com/?cdcId=dqcdc&cdcVersion=3.3&docsetId=dqgs&docsetTarget=home.htm&locale=en.

CONTACT INFORMATION

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

Brian Rineer
100 SAS Campus Drive
Cary, NC 27513
SAS Institute Inc.
+1-919-677-8000
Brian.Rineer@sas.com
http://www.sas.com