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:

![Figure 1. Data Records before Data Preparation](image)
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.

![Figure 2. SAS Data Explorer: Details Tab](image)

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.

![Figure 3. SAS Data Explorer: Sample Data Tab](image)
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](image)

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.

![Figure 5. SAS Data Explorer: Profile Tab](image)
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.

![Image of data profile report](image)

**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.
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.
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".
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.
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.
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.
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.

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.
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.
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.
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.

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.
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](https://go.documentation.sas.com/?cdcId=dprepcdc&cdcVersion=2.2&docsetId=dprepgs&docsetTarget=home.htm&locale=en)
CONTACT INFORMATION

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

Mary Kathryn Queen
SAS Institute Inc.
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.