Paper PO-17

Using Dictionary Tables to Explore SAS® Datasets

Phillip Julian, Bank of America, Charlotte, NC

ABSTRACT

Data profiling is an essential task for data management, data warehousing, and exploring SAS® datasets. TDWI (http://tdwi.org) extends the usual definition of data profiling to include data exploration. This paper presents two SAS programs, Data_Explorer and Data_Profiler, that implement the TDWI definition. These SAS programs are low-cost, free solutions for data exploration and data profiling. Data_Explorer searchs for all SAS datasets, and gathers essential dataset and file attributes into a single report. Data_Profiler summarizes the values of any SAS dataset in a generic manner, which eliminates the need for custom SQL queries and custome programs to summarize what a dataset contains. These programs have been used in banking and state government. They should also be useful in the pharmaceutical industry for validating SAS datasets and managing data repositories.

OVERVIEW

- Motivation for profiling
- Definition of data profiling
- SAS program suitability for data profiling
- Meeting the needs of data profiling according to TDWI
- Four data profiling practice areas
- Ten best practices in data profiling
- Overview of the SAS programs
- Reports and datasets
- Programming details
- References and contact information

MOTIVATION

Initial motivation → Solve a problem

- Discover what SAS datasets are available.
- Find out what the SAS datasets contain.
- Research wasn't needed; the problem defined the needs.
    - Some R&D work was done to code and test the SAS programs.

Current motivation → Meet data profiling needs

- Was this an excellent solution for data profiling?
- Research was required to find a good standard to meet.
    - TDWI (The Data Warehousing Institute) is a vendor-neutral source.

DEFINITION OF DATA PROFILING

From Philip Russom: Raising the Bar for Data Profiling:

"Data profiling is the process of examining the data available in an existing data source (e.g., a database or file) and collecting statistics and information about that data. The purpose of these statistics may be to find out whether existing data can easily be used for other purposes, provide metrics relevant to data quality and standards, assess the risk involved in integrating data for new applications, and assess whether meta data accurately describes the actual values in the source database."

FOUR PRACTICE AREAS OF DATA PROFILING

Data profiling: Create a data inventory with profiles.

- Enabled by dataset_profiler.sas, requiring only a dataset name.

Data discovery: Discover new and unknown data sources.

- Enabled by dataset_explorer.sas, requiring only a directory tree name

Data monitoring: Re-profile and discover what has changed.

- Compare SAS datasets from successive runs of dataset_explorer.sas.

Collaborative profiling: Business people add meaning to the columns.

- Add comments to SAS dataset content reports from dataset_explorer.sas.

‘

SAS Program Suitability for Data Profiling

Best Practice	Programs	Required	Suitability
Just do it!	Dataset_Profiler	Dataset location	Quick and easy to do
Profile data thoroughly	Dataset_Profiler	Dataset location	Many statistics, plus owner, permissions, and key column analysis
Produce more through data profiles	Dataset_Profiler Dataset_Explorer	Dataset location, directory trees	Many statistics, plus owner, permissions, and key column analysis
Discover and profile new data sources	Dataset_Explorer	Directory trees	Data information can be compared to find changes in structure and content
Profile data across multiple IT systems	Dataset_Profiler Dataset_Explorer	Dataset, directory, database	May require SAS In-Database technology to explore databases
Map data as you discover and profile it			No -- Programs don’t discover the foreign keys, but you can discover possible keys
Re-profile data as it evolves	Dataset_Profiler Dataset_Explorer	Directory trees	Find it with Explorer, and then profile when it changes
Re-profile data daily via data monitoring	Dataset_Explorer	Directory trees	Quick and easy to find changes
Collaborate through data profiles			No – Automation is available, but collaboration is non-programming task
Support many practices with data profiling, discovery, and monitoring			No – Integration with BI, DW, DI, DQ, and MDM are outside the scope of these programs

OVERVIEW OF THE SAS PROGRAMS

Dataset_Explorer.sas finds all SAS programs in any number of directory trees.

- Returns Excel and CSV files of all tables, directory, permissions, ownership, modification date, and attributes of every column.
- Excel data be filtered to discover suitable datasets and foreign keys.
- A file of SAS libname definitions facilitates deeper data explorations.
- See details in my paper, “Using Dictionary Tables to Explore SAS Datasets.”

Dataset_Profiler.sas analyzes uniqueness, missing values, and miscoded values, and gives detailed statistics if a column is eligible as a report variable.

- Two-pass algorithm, which is efficient even for very large data.
- Provides an enhanced contents listing, with counts of missing, non-missing, and unique values, plus percentage of same. The “Stats” column decides whether a column is suitable for reporting, which is defined by the heuristic, Unique values <300 and %Uniqueness <=10%.
- Provides a detailed list of every report variable with all of its possible values, plus statistics related to every variable in the dataset.
- Excel report includes the variable name, values, and very many statistics
- Suggested usage–Hide rows and columns that you don’t want to view. Then filter to see variables and statistics of interest. This process was used in a production environment todetermine whether a meaningful report could be produced from the dataset. In other words, did it have enough useful data to create a good analysis?
- Some details are in my paper, “Using Dictionary Tables to Explore SAS Datasets.”

Reports and Datasets – the Contents Report

Column	Value
Variable	Name of the variable
Count	Count of non-missing values
Filled	% of rows that are filled with data
NMiss	Number of missing values
Miss_Pct	% of rows with missing values
Unique	Number of unique data values
Unique_Pct	Unique / Filled formatted as % of Unique values filled
Unique_Pct_All	Unique / Count formatted as % of Unique values overall
Stats	'Y' if this can be a class value or a report variable; in other words, it's a discrete variable, and not a key or a continuous variable
Contents Data	SAS metadata values for data type, length, Format, InFormat, Label, and varnum

Dataset_Profiler.sas is detailed here. Dataset_Explorer.sas is detailed in my paper, Using Dictionary Tables to Explore SAS Datasets.

Reports and Datasets – the Statistics Report

Column	Value
Variable	Name of the variable
Values	Distinct values of Variable
Count	# of rows that have that Value
Stats	Various statistics from Proc means – see the list below for an example
Column #	Where this data comes from in the original dataset, in case you have a whole lot of columns and the data is hard to locate

PARTIAL COLUMN LISTING FROM A STATS DATASET CREATED BY DATASET_PROFILER.SAS

contact_date_Max contact_date_Mean contact_date_Min contact_date N contact_date_Nmiss contact_date_StdDev
cr_Max cr_Mean cr_Min cr_N cr_NMiss cr_StdDev
ecg_id_Max ecg_id_Mean ecg_id_Min ecg_id_N ecg_id_NMiss ecg_id_StdDev
emb_Max emb_Mean emb_Min emb_N emb_NMiss emb_StdDev
First_Prin_Bal_Max First_Prin_Bal_Mean First_Prin_Bal_Min First_Prin_Bal_N First_Prin_Bal_NMiss First_Prin_Bal_StdDev
fuba_nbr_Max fuba_nbr_Mean fuba_nbr_Min fuba_nbr_N fuba_nbr_NMiss fuba_nbr_StdDev
Last_changed_date_Max Last_changed_date_Mean Last_changed_date_Min Last_changed_date_N
Last_changed_date_NMiss Last_changed_date_StdDev

loan_no_Max loan_no_Mean loan_no_Min loan_no_N loan_no_NMiss loan_no_StdDev
ltv_Max ltv_Mean ltv_Min ltv_N ltv_NMiss ltv_StdDev

SELECTED PROGRAMMING DETAILS FOR DATASET PROFILER

1. First, define several *%let*'s at the top of the program, and choose the SAS dataset, libname, and tag for the reports.
2. If desired, define *check_keys* to look for duplicate key values.
3. Get an alphabetic-ordered list of dataset contents from the sas dictionary, and print it to an .rtf file.
4. Now comes a tricky part of the SAS code –create the sql statements that will count unique, missing, unique values, and total rows to produce the first report, which is the contents report described above.

SAS SQL statements loop through each column (Name) of the profiled dataset's metadata in SAS. SQL query lines are produced like the following phrase, and each phrase is separated by a comma for SQL. These phrases are placed into the macro variable, Count_All_Vars:

count(Name) as N_Name, count(distinct Name) as ND_Name

Here is the SAS SQL code that creates the query phrases described above:

```
proc sql noprint stimer;
select "count(" || strip(Name) || ") as N_" ||
               substr(left(Name),1,min(29, length(strip(Name)))) ||
       ", count(distinct " || strip(Name) || ") as ND_" ||
               substr(left(Name),1,min(29, length(strip(Name))))
       into :Count_All_Varsseparated by ", "
       from dictionary.columns
       where libname= upcase("&Refi_Libname") and
               memname= upcase("&Refi_DSN");
    %put Count_All_Varshas &sqlobsrows;
quit;
```

Note–Only 29 characters of a column Name are used, because a SAS variable name must not contain more than 32 characters.

The SAS macro variable is used in a SQL block that creates the dataset for the Contents Report.
Here is the SAS SQL code for that:

```
proc sql noprint stimer;
%*--Perform the dataset counts prepared above. --;
       create table key_values_0 as
       select count(*) as All_Rows, &Count_All_Vars
               from &Means_DSN;
               %put Dataset has &sqlobsrows;
quit;
```

The result of the query is a single row of data, which we will transpose into a set of variables, counts, and percents. I won't go through all of the details, but I will say that it's typical processing for Proc Transpose datasets. All you are doing is exchanging rows and columns, and then using the Transpose _Name_ variable to define the current column that is being analyzed. To fully understand how this works, you should view the dataset before and after transposition, and note that the N_ names are followed by the ND_ names, so you only have a SAS "output" statement after the ND_ names. The data step that processes the transposed data also calculates data for the Contents Report.

5. If you are checking key uniqueness, this SAS code uses a SQL paradigm to count duplicates with an **in-line view**, and then matches back to the original dataset for detail lines. Here is the code:

```
proc sql noprint stimer;
create table key_values2 as
       select a.*
               from &Means_DSNa,
               (select Loan_No, Contact_Date, count(*) as Count
               from &Means_DSN
               group by Loan_No, Contact_Date
               having count(*) > 1) b
```

```
        where a.Loan_No= b.Loan_Noand a.Contact_Date= b.Contact_Date
        order by Loan_No, Contact_Date;
    %put Dataset has &sqlobsrows;
quit;
```

Note–The **in-line view** is show in **bold**, and matching to the original dataset is shown in red. Group-By with a Having clause is a typical SQL method for finding duplicates, and the results of that query are joined with the original dataset to extract all duplicate lines.

6. More details on the SAS code for the Dataset Profiler program can be found in my SESUG 2010 paper. See the URL for that paper in the References section. Note that this presentation has two SAS programs embedded within PowerPoint. If you have any difficulty extracting that SAS code, just send a request to my email.

REFERENCES

Philip Russom, "Raising the Bar for Data Profiling", What Works in Data Integration, Volume 29, pp. 2 –5, (2010, TDWI). See www.tdwi.org or
http://tdwi.org/articles/2010/05/06/raising-the-bar-for-data-profiling.aspx?sc_lang=en

Phillip Julian, "Using Dictionary Tables to Explore SAS® Datasets", SESUG 2010 Proceedings, (2010), see
http://analytics.ncsu.edu/sesug/2010/PO23.Julian.pdf

Download this presentation with source code at
http://www.sascommunity.org/wiki/File:Using_Dictionary_Tables_to_Profile_SAS_Datasets.pptx

CONTACT INFORMATION

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

Phillip Julian
(919) 623-1309
julianp@acm.org
www.acm.org/~julianp