Paper BB-06

Using Dictionary Tables to Profile 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 searches 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, and eliminates the need for custom SQL queries to learn what the data looks like. Because the profiler uses an efficient two-pass algorithm, a brute force approach, that includes everything plus the kitchen sink, can consume fewer resources than custom SQL queries. Profiler results are more complete because you get complete categorical details for all the columns of very big datasets.

These programs have been used in banking and state government, and should be useful in the pharmaceutical industry for validating SAS datasets and managing data content and changes in large data repositories.

INTRODUCTION

Terse bullet points and annotated figures emphasize that this is a visual presentation of technical information. The two data profiling programs are complex, but the visual results are easily understood by anyone. Therefore programming details are deferred until the second half of this paper.

Results from the SAS programs are explained by a concrete example in the related poster (PO-01, SESUG 2012). Each section of the poster is discussed to show how you may use the profile reports to find and fix real issues. Analysts, report writers, and data experts may picture solutions as they see this information. With the end in sight and the anticipation of solving difficult problems, they may be more interested in seeing how these programs work.

Both SAS programs use Base SAS and SAS/Connect. You can remove the dependency on SAS/Connect by rewriting the programs as local SAS programs. Data_Explorer is written for SAS on UNIX, but it can easily be adapted to other systems.

OVERVIEW

- Results First, Then the Technical Details
    - Limitations of Metadata
    - Seeing is Believing
- Seeing the Visual Results
- The complete poster, suitable for printing
- Motivation for Data Profiling
- Definition of Data Profiling
- Features of a Good Data Profiler
    - 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

RESULTS FIRST, THEN THE TECHNICAL DETAILS

- Limitations of Metadata
  - Metadata information is hidden behind complicated GUIs.
  - Many choices come from metadata searches, and they all look good.
  - Descriptions may not have enough information about all columns.
  - Descriptions may not be clear and concise, so the truth gets lost in the overwhelming details.
  - Column descriptions may not show the complete set of values in the data.
  - Column descriptions may not be current.
  - Subject matter experts (SMEs) may be required to make a good decision.
  - SMEs may be hard to find, not available, or unknown.

- Seeing is Believing
  - When metadata is lacking, not authoritative, or not explained very well, you view the data and run custom queries to learn about the data.
  - Seeing is believing, and it may be your only recourse when you lack good metadata descriptions.
  - Even when metadata is very good, you still view the data if it's new to you.
  - Seeing is believing, even when you have good metadata.

SEEING THE VISUAL RESULTS – FEATURES AND APPLICATIONS

Figures 1 – 6 explain each part of the poster (PO-01, SESUG 2012). Figure 1 shows the features of the programs, and some possible uses of the information. The complete poster is in Figure 7, followed by four images that can be printed on letter-sized pages and taped together to create the poster.

The following excerpt shows the upper left corner of the poster. You can see that the profiler report is in Excel, and various columns have been hidden to explore counts of non-missing values for selected columns.

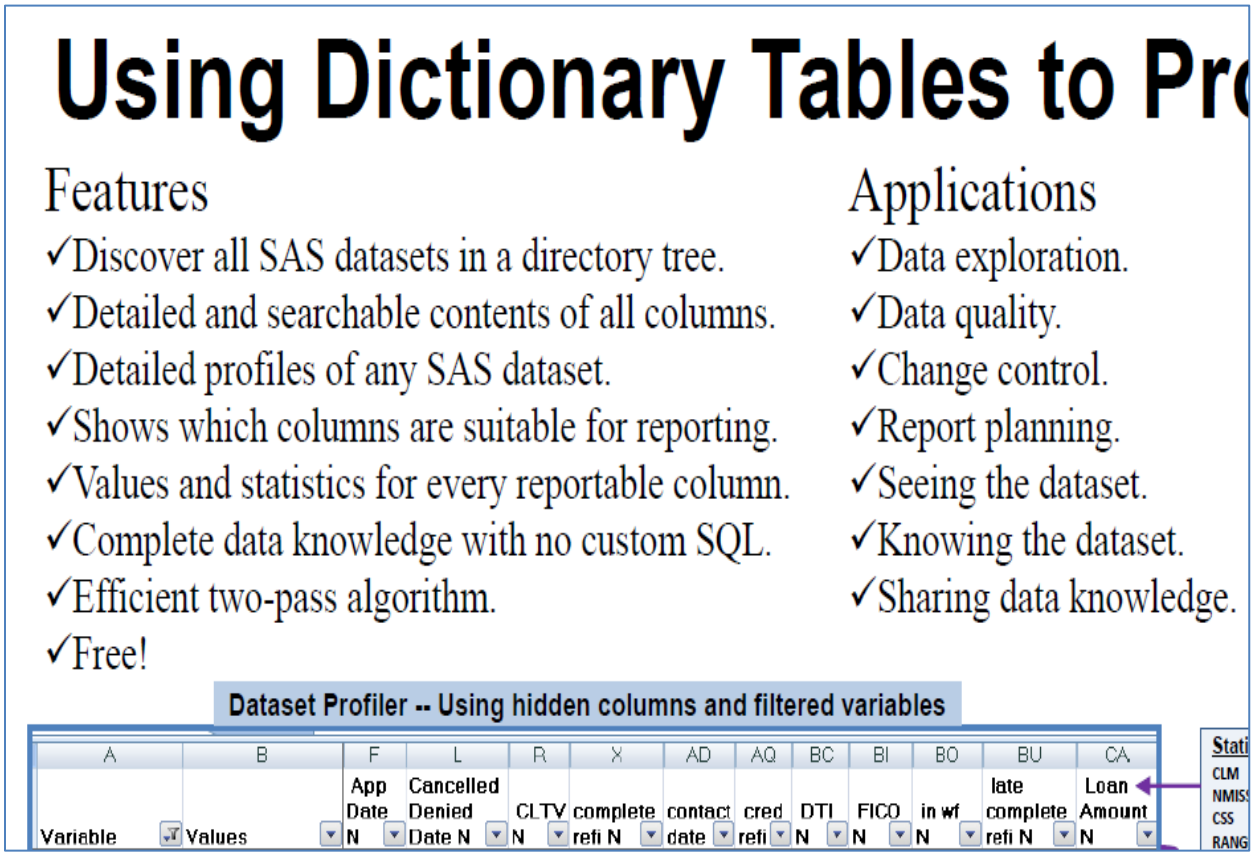


Figure 1. Seeing the Visual Results – Features and Applications

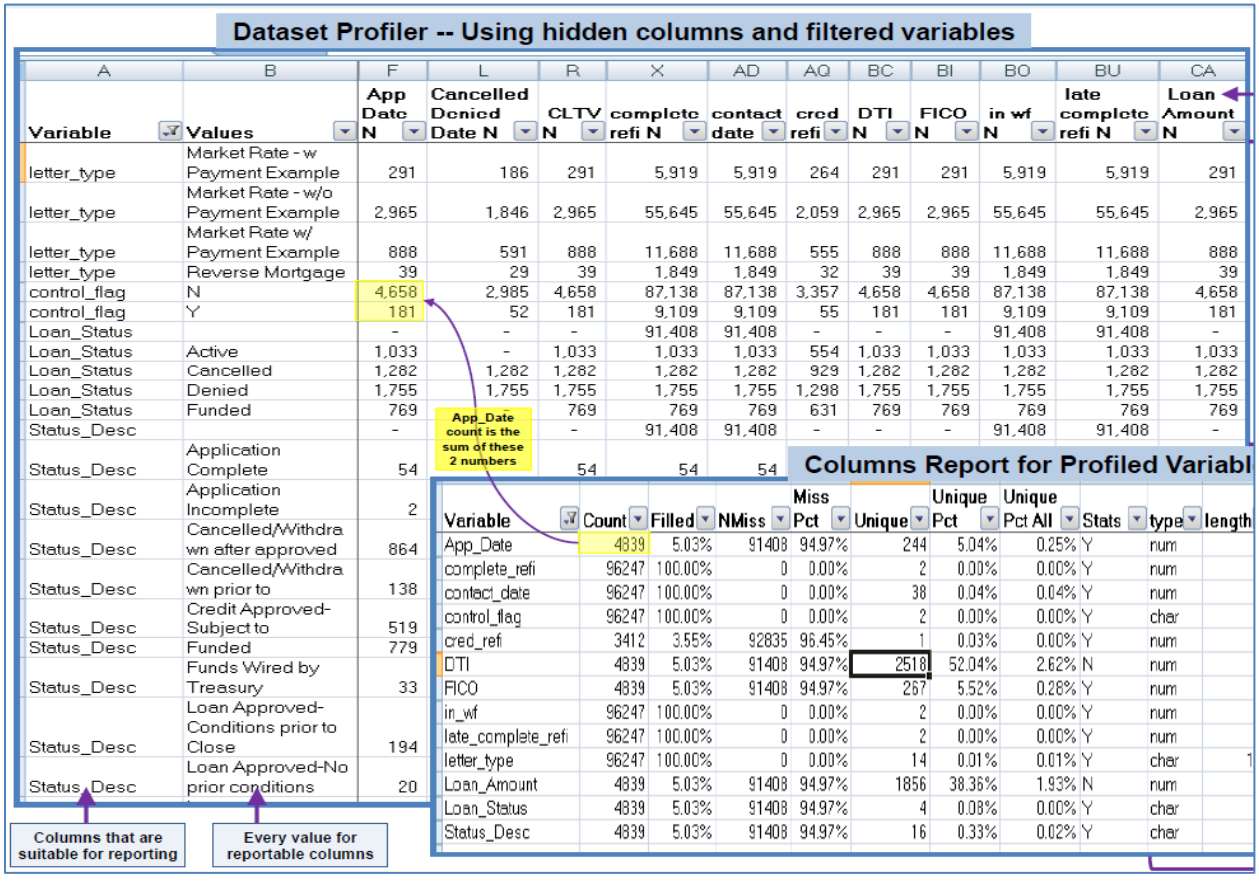
SEEING THE VISUAL RESULTS – DATASET PROFILER

The poster is highlighted and annotated with arrows and boxes, as you can see from this excerpt from the upper right corner. Starting with the boxes on the bottom left, columns A and B of the profile report are explained (*Dataset Profiler – Using hidden columns and filtered variables*):

- Column A contains column names, which are repeated for each distinct value of a column.
- Column B shows the distinct values for each column, including missing values.
- The other columns show statistics for all numeric columns when Column A has that value in Column B.
  - Almost any SAS statistic can be calculated here. Refer to Figure 3.
  - Only "reportable" columns are shown in Column A.
  - Reportable variables have no more than 300 distinct values, which you can adjust in the program.

The profile report is overlaid on lower right with an enhanced content report on the dataset (*Columns Report for Profiled Variables*). Notice the rich set of statistics for each variable.

- The Stats column in the report is Y if it is a reportable variable, and N otherwise.
- All columns in the report are defined in Figure 4.
- Note the highlighted relationship between the App_Date variable and the App_Date_N column in the Dataset Profiler report. The 4839 non-missing values of App_Date are either Y or N in the Dataset Profiler report.
- Note that analytical groups of variables are defined by their Count values. Six variables have 4839 non-missing values, and cred_refi is close at 3412. These variables can be statistically analyzed as a group. Other variables have no missing values, and they represent another group of variables for related analysis.



A	B	F	L	R	X	AD	AQ	BC	BI	BO	BU	CA
Variable	Values	App Date N	Cancelled Denied Date N	CLTV N	complete refi N	contact date	cred refi	DTI N	FICO N	in wf N	late complete refi N	Loan Amount N
letter_type	Market Rate - w Payment Example	291	186	291	5,919	5,919	264	291	291	5,919	5,919	291
letter_type	Market Rate - w/o Payment Example	2,965	1,846	2,965	55,645	55,645	2,059	2,965	2,965	55,645	55,645	2,965
letter_type	Market Rate w/ Payment Example	888	591	888	11,688	11,688	555	888	888	11,688	11,688	888
letter_type	Reverse Mortgage	39	29	39	1,849	1,849	32	39	39	1,849	1,849	39
control_flag	N	4,658	2,985	4,658	87,138	87,138	3,357	4,658	4,658	87,138	87,138	4,658
control_flag	Y	181	52	181	9,109	9,109	55	181	181	9,109	9,109	181
Loan_Status		-	-	-	91,408	91,408	-	-	-	91,408	91,408	-
Loan_Status	Active	1,033	-	1,033	1,033	1,033	554	1,033	1,033	1,033	1,033	1,033
Loan_Status	Cancelled	1,282	1,282	1,282	1,282	1,282	929	1,282	1,282	1,282	1,282	1,282
Loan_Status	Denied	1,755	1,755	1,755	1,755	1,755	1,298	1,755	1,755	1,755	1,755	1,755
Loan_Status	Funded	769	-	769	769	769	631	769	769	769	769	769
Status_Desc		-	-	-	91,408	91,408	-	-	-	91,408	91,408	-
Status_Desc	Application Complete	54		54	54	54						
Status_Desc	Application Incomplete	2										
Status_Desc	Cancelled/Withdrawn after approved	864										
Status_Desc	Cancelled/Withdrawn prior to	138										
Status_Desc	Credit Approved-Subject to	519										
Status_Desc	Funded	779										
Status_Desc	Funds Wired by Treasury	33										
Status_Desc	Loan Approved-Conditions prior to Close	194										
Status_Desc	Loan Approved-No prior conditions	20										

Variable	Count	Filled	NMiss	Miss Pct	Unique	Unique Pct	Unique Pct All	Stats	type	length
App_Date	4839	5.03%	91408	94.97%	244	5.04%	0.25%	Y	num	
complete_refi	96247	100.00%	0	0.00%	2	0.00%	0.00%	Y	num	
contact_date	96247	100.00%	0	0.00%	38	0.04%	0.04%	Y	num	
control_flag	96247	100.00%	0	0.00%	2	0.00%	0.00%	Y	char	
cred_refi	3412	3.55%	92835	96.45%	1	0.03%	0.00%	Y	num	
DTI	4839	5.03%	91408	94.97%	2518	52.04%	2.62%	N	num	
FICO	4839	5.03%	91408	94.97%	267	5.52%	0.28%	Y	num	
in_wf	96247	100.00%	0	0.00%	2	0.00%	0.00%	Y	num	
late_complete_refi	96247	100.00%	0	0.00%	2	0.00%	0.00%	Y	num	
letter_type	96247	100.00%	0	0.00%	14	0.01%	0.01%	Y	char	1
Loan_Amount	4839	5.03%	91408	94.97%	1856	38.36%	1.93%	N	num	
Loan_Status	4839	5.03%	91408	94.97%	4	0.08%	0.00%	Y	char	
Status_Desc	4839	5.03%	91408	94.97%	16	0.33%	0.02%	Y	char	

Figure 2. Seeing the Visual Results – Features and Applications

SEEING THE VISUAL RESULTS – PROFILER STATISTICS

A wide variety of statistics are available for the profiler report, which uses Proc Means. Statistics may be selected by modifying the Dataset_Profiler program. The example report shows the N statistic, the number of non-missing values of a variable.

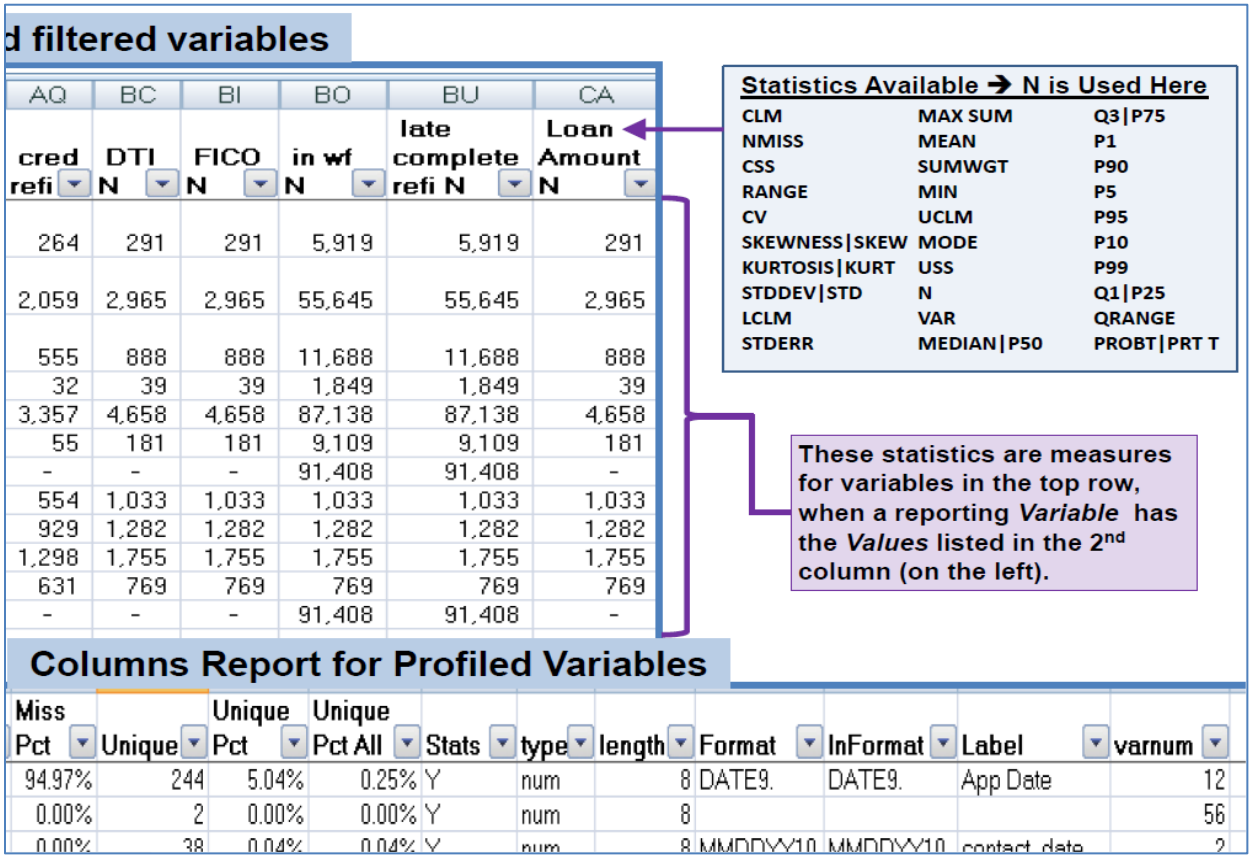
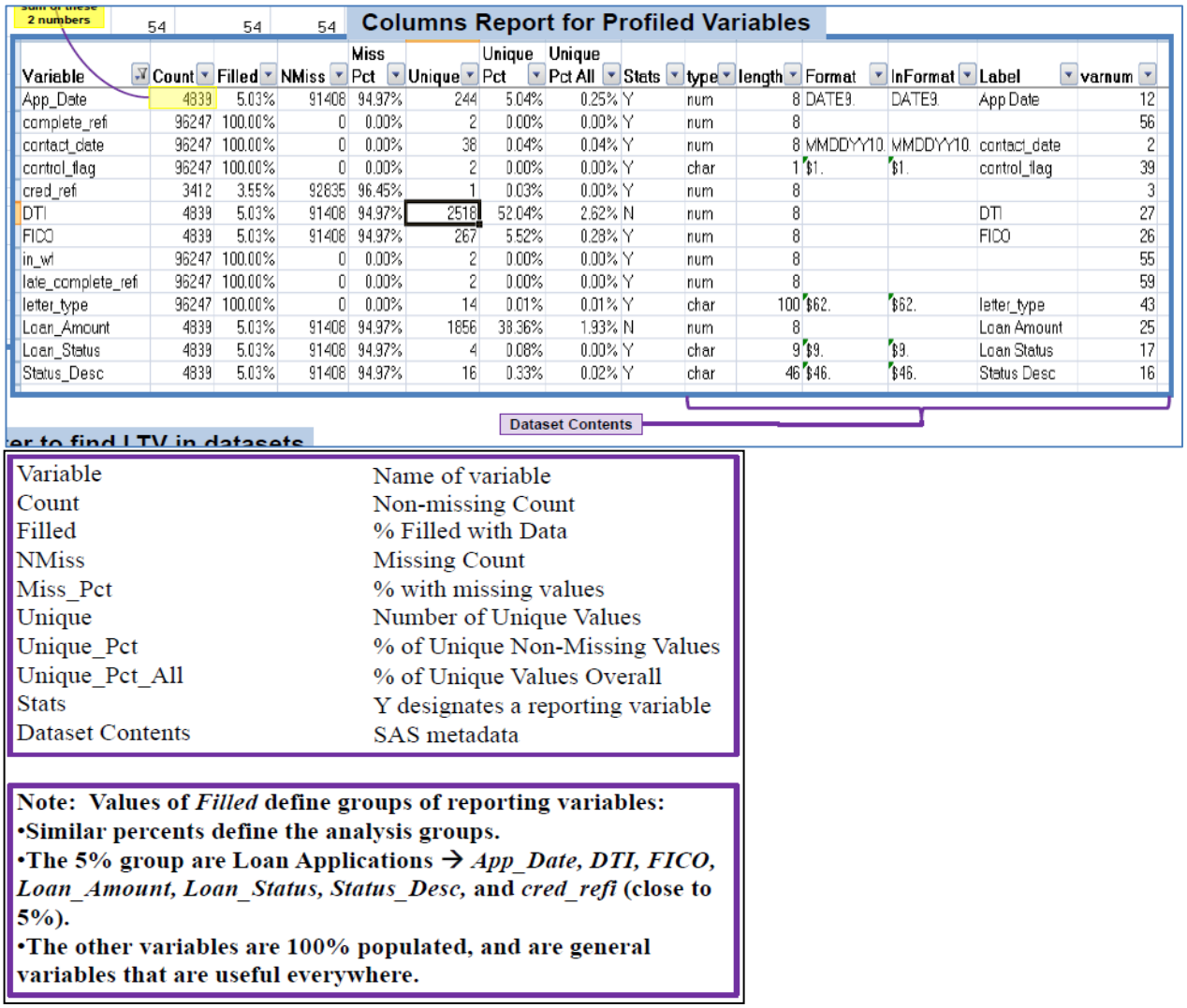


Figure 3. Seeing the Visual Results – Profiler Statistics

SEEING THE VISUAL RESULTS – COLUMN REPORT DEFINITIONS

Definitions for columns of the *Columns Report for Profiled Variables* are shown in the visual below.

- The upper box shows the column definitions for the report.
- The lower box describes the two analytical groupings in the report, based upon similar counts of non-missing values in two groups of columns.



Variable	Count	Filled	NMiss	Miss Pct	Unique	Unique Pct	Unique Pct All	Stats	type	length	Format	InFormat	Label	varnum
App_Date	4839	5.03%	91408	94.97%	244	5.04%	0.25%	Y	num	8	DATE9.	DATE9.	App Date	12
complete_refi	96247	100.00%	0	0.00%	2	0.00%	0.00%	Y	num	8				56
contact_date	96247	100.00%	0	0.00%	38	0.04%	0.04%	Y	num	8	MMDDYY10.	MMDDYY10.	contact_date	2
control_flag	96247	100.00%	0	0.00%	2	0.00%	0.00%	Y	char	1	$1.	$1.	control_flag	39
cred_refi	3412	3.55%	92835	96.45%	1	0.03%	0.00%	Y	num	8				3
DTI	4839	5.03%	91408	94.97%	2518	52.04%	2.62%	N	num	8			DTI	27
FICO	4839	5.03%	91408	94.97%	267	5.52%	0.28%	Y	num	8			FICO	26
in_wi	96247	100.00%	0	0.00%	2	0.00%	0.00%	Y	num	8				55
late_complete_refi	96247	100.00%	0	0.00%	2	0.00%	0.00%	Y	num	8				59
letter_type	96247	100.00%	0	0.00%	14	0.01%	0.01%	Y	char	100	$62.	$62.	letter_type	43
Loan_Amount	4839	5.03%	91408	94.97%	1856	38.36%	1.93%	N	num	8			Loan Amount	25
Loan_Status	4839	5.03%	91408	94.97%	4	0.08%	0.00%	Y	char	9	$9.	$9.	Loan Status	17
Status_Desc	4839	5.03%	91408	94.97%	16	0.33%	0.02%	Y	char	46	$46.	$46.	Status Desc	16

Figure 4. Seeing the Visual Results – Column Report Definitions

SEEING THE VISUAL RESULTS – DATA QUALITY ANALYSIS

The dataset explorer report can be filtered to display anomalies in column attributes. In this example, LTV is filtered in the upper panel, and results are shown in the lower panel. LTV should be numeric, but variables in some tables are defined as character (see the purple box in the image below). This inconsistency can cause program failure or inaccurate results. Also note that the format of LTV is not uniform, and that may affect the report view of the data.

Note that the dataset name and directory are displayed. This data can be used to create a program to make corrections to the data.

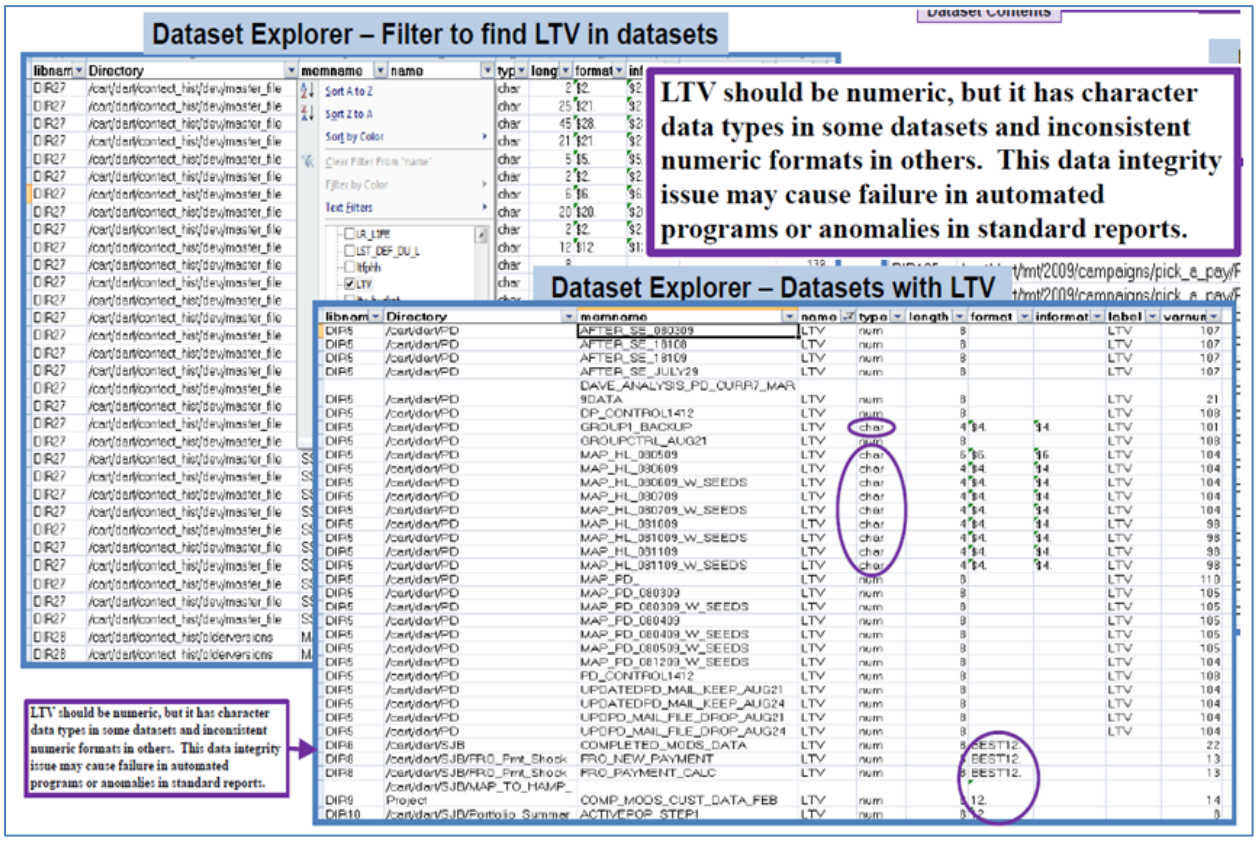


Figure 5. Seeing the Visual Results – Data Quality Analysis

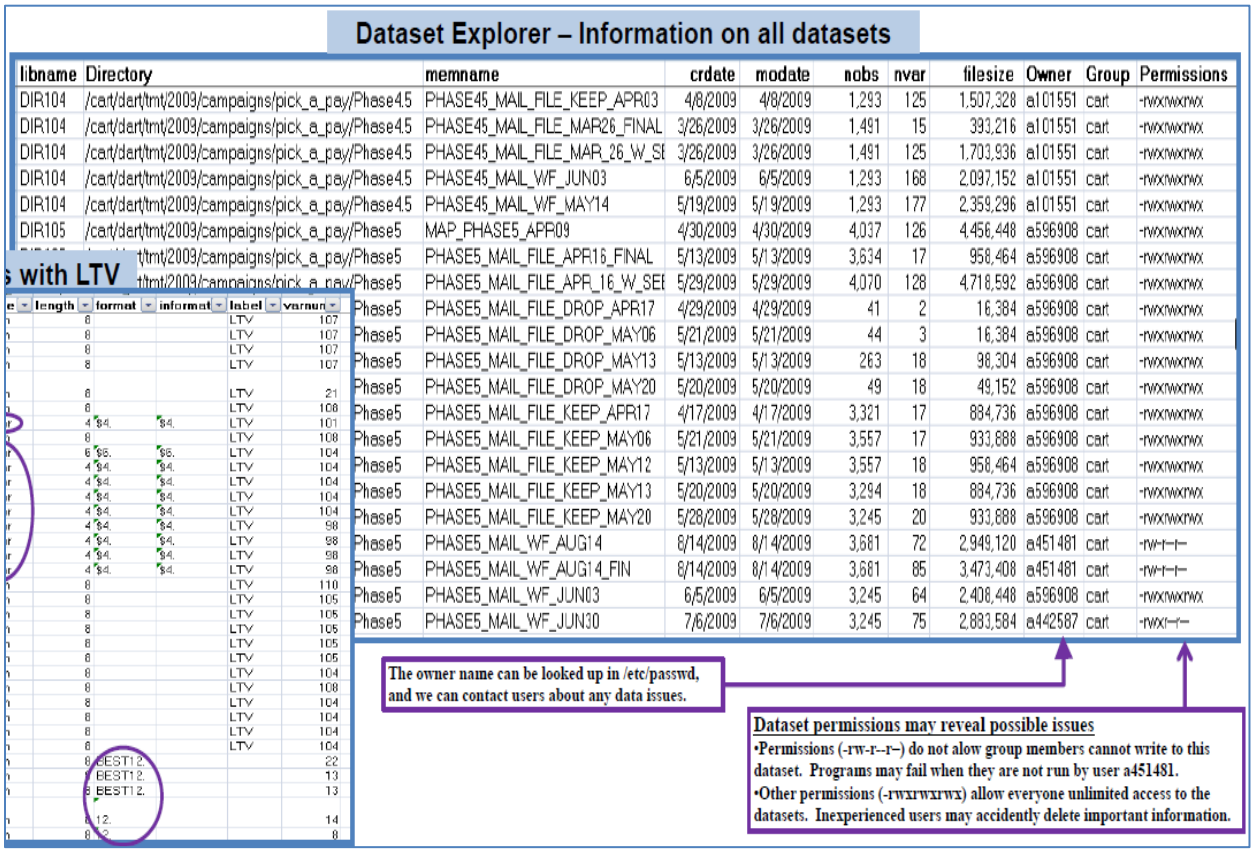
SEEING THE VISUAL RESULTS – EXPLORER COLUMNS AND PERMISSIONS

The dataset explorer is an enhanced Proc Contents report that lists the owner, group, and permissions for all SAS datasets. These UNIX attributes can cause problems for programs or users that need to access datasets or create new ones.

Notice that very many datasets have permissions that are too loose. This caused problems when datasets were deleted and recreated with 20 observations by a novice SAS user. The report would show Owner names so you can locate the errant user and give them more training on SAS.

Other datasets had permissions that did not allow group members to recreate the dataset. This caused program failure when the marketing campaign was run by another person.

In all these cases, you can locate the owner by the Owner column, find their email in /etc/passwd, and notify them about the problem with permissions.



libname	Directory	memname	crdate	modate	nobs	nvar	filesize	Owner	Group	Permissions
DIR104	/cart/dart/tmt/2003/campaigns/pick_a_pay/Phase4.5	PHASE45_MAIL_FILE_KEEP_APR03	4/8/2009	4/8/2009	1,293	125	1,507,328	a101551	cart	-rwxrwxrwx
DIR104	/cart/dart/tmt/2003/campaigns/pick_a_pay/Phase4.5	PHASE45_MAIL_FILE_MAR26_FINAL	3/26/2009	3/26/2009	1,491	15	393,216	a101551	cart	-rwxrwxrwx
DIR104	/cart/dart/tmt/2003/campaigns/pick_a_pay/Phase4.5	PHASE45_MAIL_FILE_MAR_26_W_SE	3/26/2009	3/26/2009	1,491	125	1,703,936	a101551	cart	-rwxrwxrwx
DIR104	/cart/dart/tmt/2003/campaigns/pick_a_pay/Phase4.5	PHASE45_MAIL_WF_JUN03	6/5/2009	6/5/2009	1,293	168	2,097,152	a101551	cart	-rwxrwxrwx
DIR104	/cart/dart/tmt/2003/campaigns/pick_a_pay/Phase4.5	PHASE45_MAIL_WF_MAY14	5/19/2009	5/19/2009	1,293	177	2,359,296	a101551	cart	-rwxrwxrwx
DIR105	/cart/dart/tmt/2003/campaigns/pick_a_pay/Phase5	MAP_PHASE5_APR09	4/30/2009	4/30/2009	4,037	126	4,456,448	a596908	cart	-rwxrwxrwx
	t/tmt/2003/campaigns/pick_a_pay/Phase5	PHASE5_MAIL_FILE_APR16_FINAL	5/13/2009	5/13/2009	3,634	17	958,464	a596908	cart	-rwxrwxrwx
	t/tmt/2003/campaigns/pick_a_pay/Phase5	PHASE5_MAIL_FILE_APR_16_W_SEI	5/29/2009	5/29/2009	4,070	128	4,718,592	a596908	cart	-rwxrwxrwx
	Phase5	PHASE5_MAIL_FILE_DROP_APR17	4/29/2009	4/29/2009	41	2	16,384	a596908	cart	-rwxrwxrwx
	Phase5	PHASE5_MAIL_FILE_DROP_MAY06	5/21/2009	5/21/2009	44	3	16,384	a596908	cart	-rwxrwxrwx
	Phase5	PHASE5_MAIL_FILE_DROP_MAY13	5/13/2009	5/13/2009	263	18	98,304	a596908	cart	-rwxrwxrwx
	Phase5	PHASE5_MAIL_FILE_DROP_MAY20	5/20/2009	5/20/2009	49	18	49,152	a596908	cart	-rwxrwxrwx
	Phase5	PHASE5_MAIL_FILE_KEEP_APR17	4/17/2009	4/17/2009	3,321	17	884,736	a596908	cart	-rwxrwxrwx
	Phase5	PHASE5_MAIL_FILE_KEEP_MAY06	5/21/2009	5/21/2009	3,557	17	933,888	a596908	cart	-rwxrwxrwx
	Phase5	PHASE5_MAIL_FILE_KEEP_MAY12	5/13/2009	5/13/2009	3,557	18	958,464	a596908	cart	-rwxrwxrwx
	Phase5	PHASE5_MAIL_FILE_KEEP_MAY13	5/20/2009	5/20/2009	3,294	18	884,736	a596908	cart	-rwxrwxrwx
	Phase5	PHASE5_MAIL_FILE_KEEP_MAY20	5/28/2009	5/28/2009	3,245	20	933,888	a596908	cart	-rwxrwxrwx
	Phase5	PHASE5_MAIL_WF_AUG14	8/14/2009	8/14/2009	3,681	72	2,949,120	a451481	cart	-rw-r--r--
	Phase5	PHASE5_MAIL_WF_AUG14_FIN	8/14/2009	8/14/2009	3,681	85	3,473,408	a451481	cart	-rw-r--r--
	Phase5	PHASE5_MAIL_WF_JUN03	6/5/2009	6/5/2009	3,245	64	2,408,448	a596908	cart	-rwxrwxrwx
	Phase5	PHASE5_MAIL_WF_JUN30	7/6/2009	7/6/2009	3,245	75	2,883,584	a442587	cart	-rwxr--r--

Figure 6. Seeing the Visual Results – Explorer Columns and Permissions

SEEING THE VISUAL RESULTS – THE COMPLETE POSTER

The complete poster appears on the next page. The visual resolution of the complete poster is too coarse to create a good print. Therefore, four pages of higher resolution images are included for printing and taping together to create a better version of the complete poster.

Here is the complete poster on a single page. The next 4 pages contain images suitable for printing and taping.

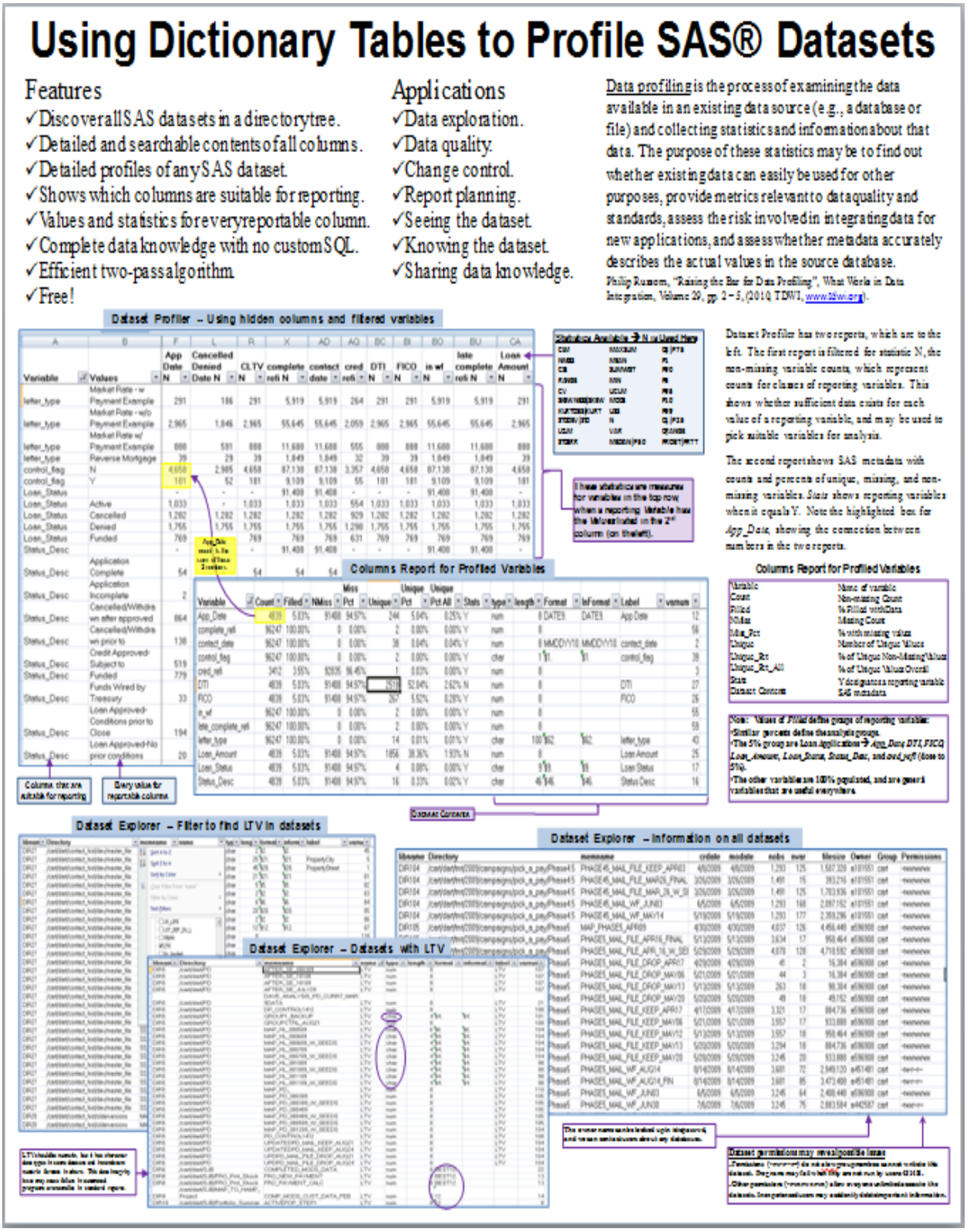
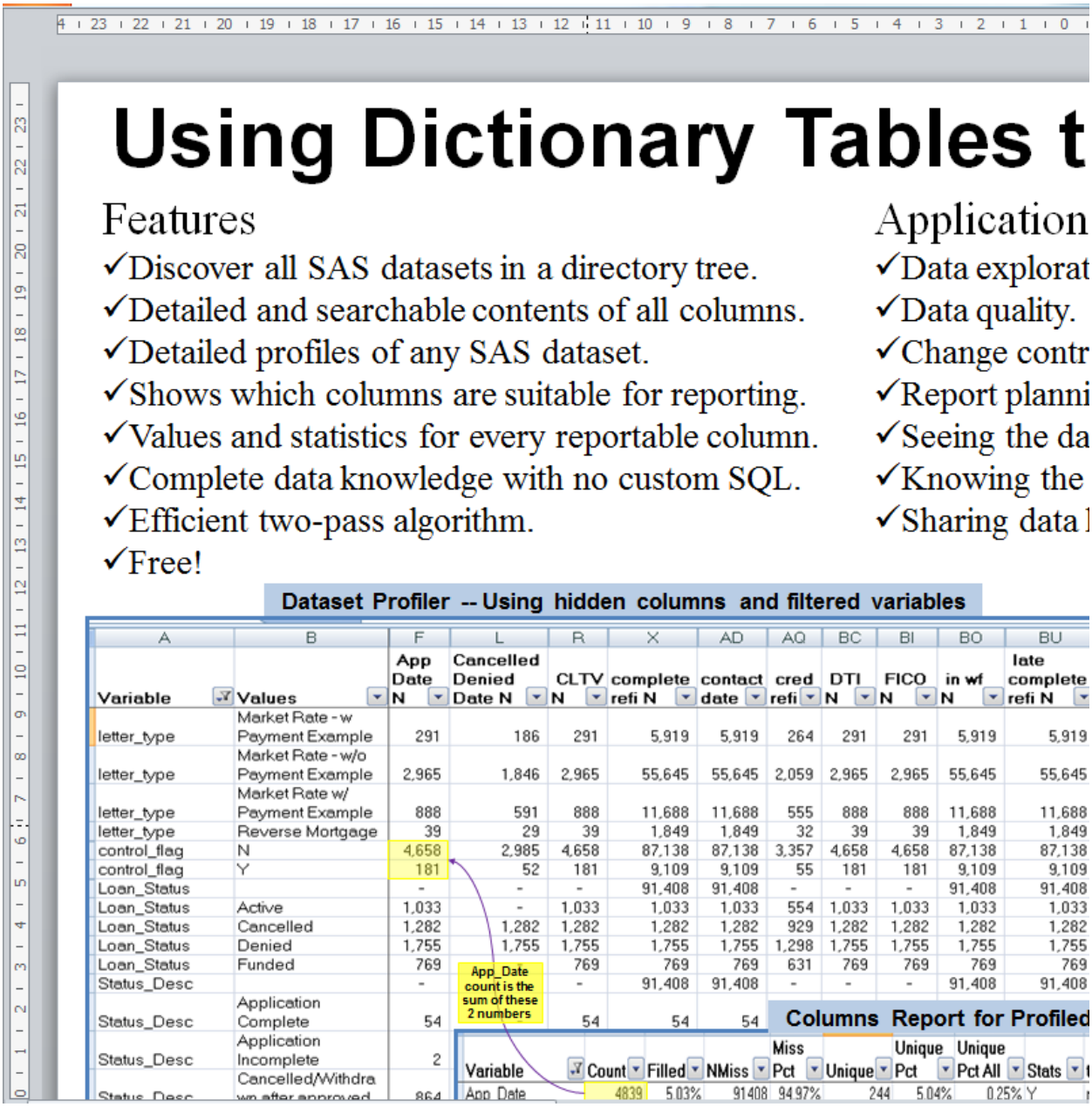


Figure 7. Seeing the Visual Results – The Complete Poster



Using Dictionary Tables t

Features

- ✓Discover all SAS datasets in a directory tree.
- ✓Detailed and searchable contents of all columns.
- ✓Detailed profiles of any SAS dataset.
- ✓Shows which columns are suitable for reporting.
- ✓Values and statistics for every reportable column.
- ✓Complete data knowledge with no custom SQL.
- ✓Efficient two-pass algorithm.
- ✓Free!

Application

- ✓Data explorat
- ✓Data quality.
- ✓Change contr
- ✓Report planni
- ✓Seeing the da
- ✓Knowing the
- ✓Sharing data

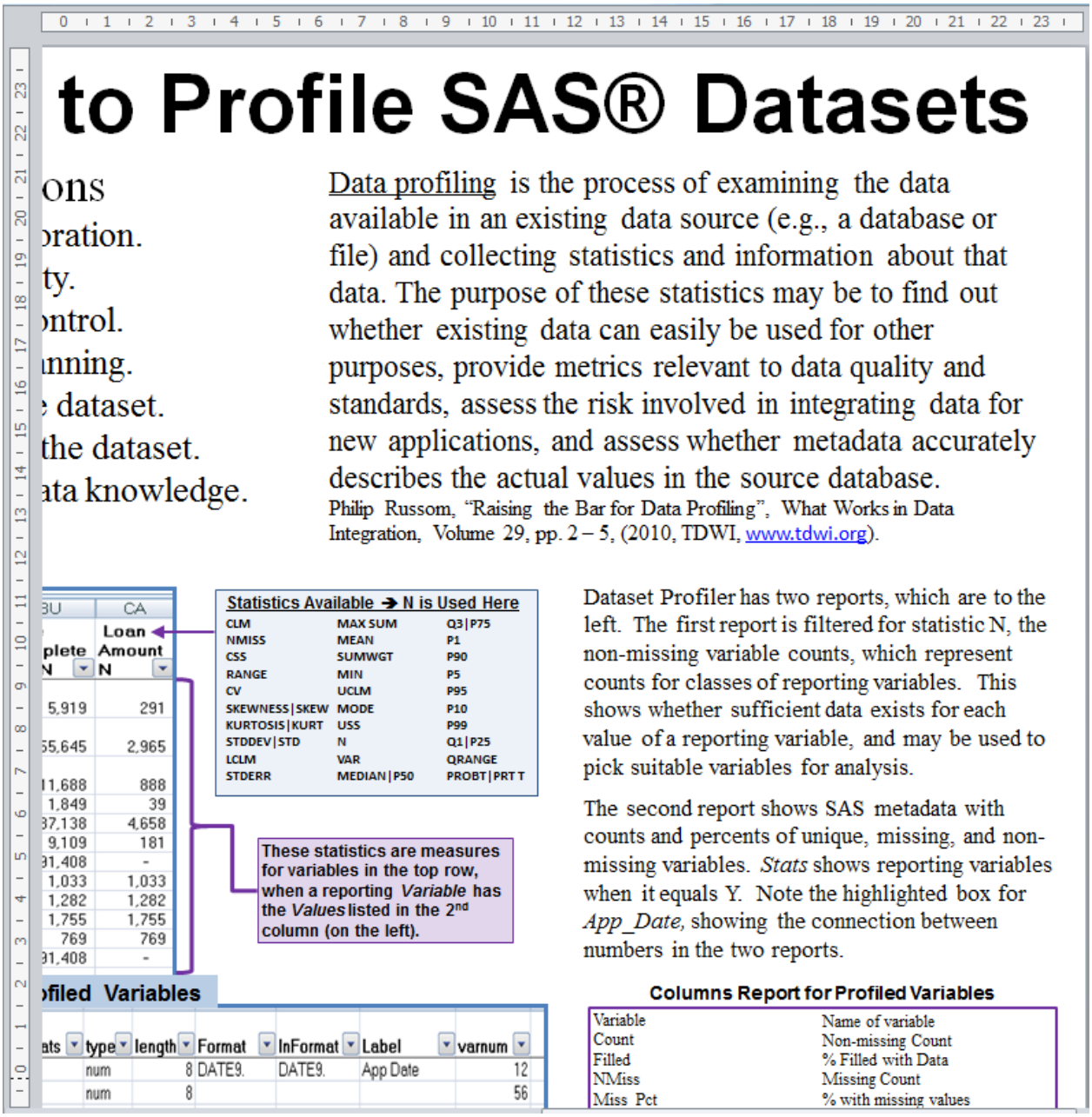
Dataset Profiler -- Using hidden columns and filtered variables

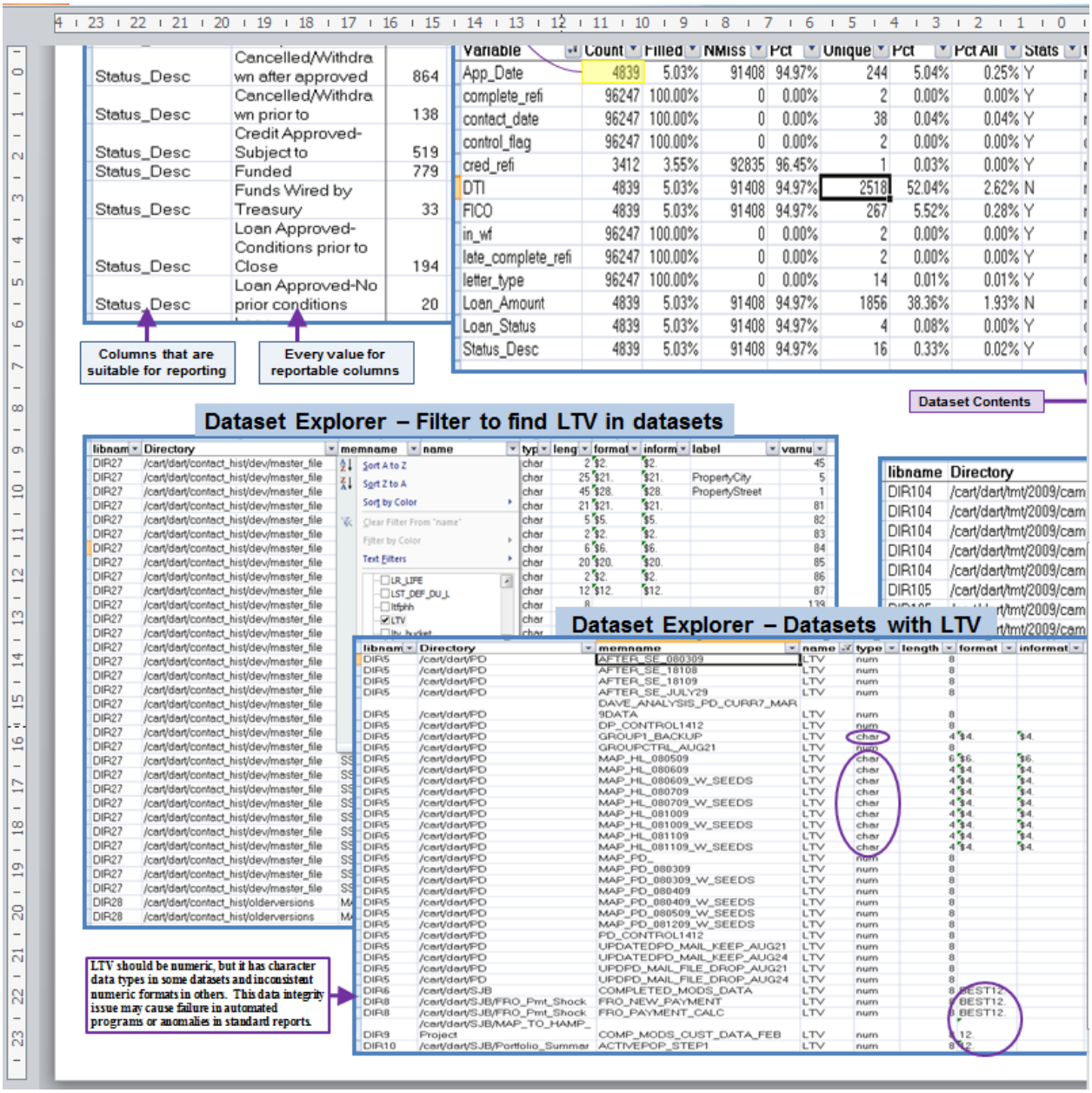
A	B	F	L	R	X	AD	AQ	BC	BI	BO	BU
Variable	Values	App Date N	Cancelled Denied Date N	CLTV N	complete refi N	contact date	cred refi	DTI N	FICO N	in wf	late complete refi N
letter_type	Market Rate - w Payment Example	291	186	291	5,919	5,919	264	291	291	5,919	5,919
letter_type	Market Rate - w/o Payment Example	2,965	1,846	2,965	55,645	55,645	2,059	2,965	2,965	55,645	55,645
letter_type	Market Rate w/ Payment Example	888	591	888	11,688	11,688	555	888	888	11,688	11,688
letter_type	Reverse Mortgage	39	29	39	1,849	1,849	32	39	39	1,849	1,849
control_flag	N	4,658	2,985	4,658	87,138	87,138	3,357	4,658	4,658	87,138	87,138
control_flag	Y	181	52	181	9,109	9,109	55	181	181	9,109	9,109
Loan_Status		-	-	-	91,408	91,408	-	-	-	91,408	91,408
Loan_Status	Active	1,033	-	1,033	1,033	1,033	554	1,033	1,033	1,033	1,033
Loan_Status	Cancelled	1,282	1,282	1,282	1,282	1,282	929	1,282	1,282	1,282	1,282
Loan_Status	Denied	1,755	1,755	1,755	1,755	1,755	1,298	1,755	1,755	1,755	1,755
Loan_Status	Funded	769		769	769	769	631	769	769	769	769
Status_Desc		-		-	91,408	91,408	-	-	-	91,408	91,408
Status_Desc	Application Complete	54		54	54	54					
Status_Desc	Application Incomplete	2									
Status_Desc	Cancelled/Withdra wn after approved	864									

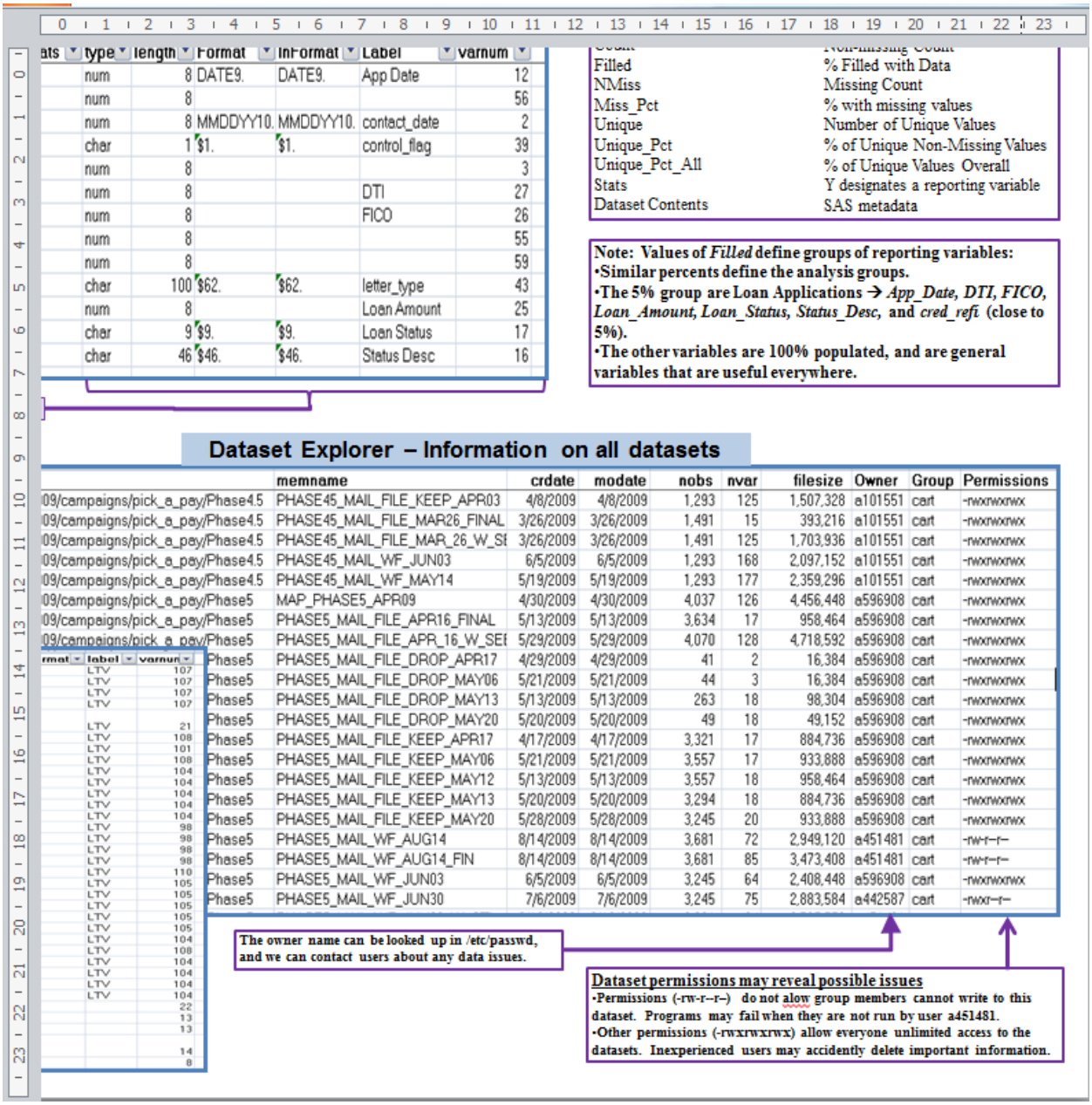
App_Date count is the sum of these 2 numbers

Columns Report for Profiled

Variable	Count	Filled	NMiss	Miss Pct	Unique	Unique Pct	Unique Pct All	Stats
App_Date	4839	5.03%	91408	94.97%	244	5.04%	0.25%	Y


to Profile SAS® Datasets
ons
oration.
ty.
ontrol.
nning.
dataset.
the dataset.
ata knowledge.
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 metadata accurately describes the actual values in the source database.
Philip Russom, "Raising the Bar for Data Profiling", What Works in Data Integration, Volume 29, pp. 2 – 5, (2010, TDWI, www.tdwi.org).
Statistics Available → N is Used Here
CLM MAX SUM Q3|P75
NMISS MEAN P1
CSS SUMWGT P90
RANGE MIN P5
CV UCLM P95
SKEWNESS|SKEW MODE P10
KURTOSIS|KURT USS P99
STDDEV|STD N Q1|P25
LCLM VAR QRANGE
STDERR MEDIAN|P50 PROBT|PRT T
These statistics are measures for variables in the top row, when a reporting Variable has the Values listed in the 2nd column (on the left).
Loan Amount N
5,919 291
55,645 2,965
11,688 888
1,849 39
37,138 4,658
9,109 181
31,408 -
1,033 1,033
1,282 1,282
1,755 1,755
769 769
31,408 -
ofiled Variables
type length Format InFormat Label varnum
num 8 DATE9. DATE9. App Date 12
num 8 56
Dataset Profiler has two reports, which are to the left. The first report is filtered for statistic N, the non-missing variable counts, which represent counts for classes of reporting variables. This shows whether sufficient data exists for each value of a reporting variable, and may be used to pick suitable variables for analysis.
The second report shows SAS metadata with counts and percents of unique, missing, and non-missing variables. Stats shows reporting variables when it equals Y. Note the highlighted box for App_Date, showing the connection between numbers in the two reports.
Columns Report for Profiled Variables
Variable Name of variable
Count Non-missing Count
Filled % Filled with Data
NMiss Missing Count
Miss_Pct % with missing values


Status_Desc Cancelled/Withdrawn after approved 864
Status_Desc Cancelled/Withdrawn prior to 138
Status_Desc Credit Approved-Subject to 519
Status_Desc Funded 779
Status_Desc Funds Wired by Treasury 33
Status_Desc Loan Approved-Conditions prior to Close 194
Status_Desc Loan Approved-No prior conditions 20
Columns that are suitable for reporting
Every value for reportable columns
Variable Count Filled NMiss Pct Unique Pct Pct All Stats
App_Date 4839 5.03% 91408 94.97% 244 5.04% 0.25% Y
complete_refi 96247 100.00% 0 0.00% 2 0.00% 0.00% Y
contact_date 96247 100.00% 0 0.00% 38 0.04% 0.04% Y
control_flag 96247 100.00% 0 0.00% 2 0.00% 0.00% Y
cred_refi 3412 3.55% 92835 96.45% 1 0.03% 0.00% Y
DTI 4839 5.03% 91408 94.97% 2518 52.04% 2.62% N
FICO 4839 5.03% 91408 94.97% 267 5.52% 0.28% Y
in_wf 96247 100.00% 0 0.00% 2 0.00% 0.00% Y
late_complete_refi 96247 100.00% 0 0.00% 2 0.00% 0.00% Y
letter_type 96247 100.00% 0 0.00% 14 0.01% 0.01% Y
Loan_Amount 4839 5.03% 91408 94.97% 1856 38.36% 1.93% N
Loan_Status 4839 5.03% 91408 94.97% 4 0.08% 0.00% Y
Status_Desc 4839 5.03% 91408 94.97% 16 0.33% 0.02% Y
Dataset Contents
Dataset Explorer – Filter to find LTV in datasets
Dataset Explorer – Datasets with LTV
LTV should be numeric, but it has character data types in some datasets and inconsistent numeric formats in others. This data integrity issue may cause failure in automated programs or anomalies in standard reports.



ats	type	length	Format	Informat	Label	varnum
	num	8	DATE9.	DATE9.	App Date	12
	num	8				56
	num	8	MMDDYY10.	MMDDYY10.	contact_date	2
	char	1	$1.	$1.	control_flag	39
	num	8				3
	num	8			DTI	27
	num	8			FICO	26
	num	8				55
	num	8				59
	char	100	$62.	$62.	letter_type	43
	num	8			Loan Amount	25
	char	9	$9.	$9.	Loan Status	17
	char	46	$46.	$46.	Status Desc	16

Filled	% Filled with Data
NMiss	Missing Count
Miss_Pct	% with missing values
Unique	Number of Unique Values
Unique_Pct	% of Unique Non-Missing Values
Unique_Pct_All	% of Unique Values Overall
Stats	Y designates a reporting variable
Dataset Contents	SAS metadata

Note: Values of *Filled* define groups of reporting variables:
•Similar percents define the analysis groups.
•The 5% group are Loan Applications → *App_Date, DTI, FICO, Loan_Amount, Loan_Status, Status_Desc,* and *cred_refi* (close to 5%).
•The other variables are 100% populated, and are general variables that are useful everywhere.

Dataset Explorer – Information on all datasets

	memname	crdate	modate	nobs	nvar	filesize	Owner	Group	Permissions
09/campaigns/pick_a_pay/Phase4.5	PHASE45_MAIL_FILE_KEEP_APR03	4/8/2009	4/8/2009	1,293	125	1,507,328	a101551	cart	-rwxrwxrwx
09/campaigns/pick_a_pay/Phase4.5	PHASE45_MAIL_FILE_MAR26_FINAL	3/26/2009	3/26/2009	1,491	15	393,216	a101551	cart	-rwxrwxrwx
09/campaigns/pick_a_pay/Phase4.5	PHASE45_MAIL_FILE_MAR_26_W_SE	3/26/2009	3/26/2009	1,491	125	1,703,936	a101551	cart	-rwxrwxrwx
09/campaigns/pick_a_pay/Phase4.5	PHASE45_MAIL_WF_JUN03	6/5/2009	6/5/2009	1,293	168	2,097,152	a101551	cart	-rwxrwxrwx
09/campaigns/pick_a_pay/Phase4.5	PHASE45_MAIL_WF_MAY14	5/19/2009	5/19/2009	1,293	177	2,359,296	a101551	cart	-rwxrwxrwx
09/campaigns/pick_a_pay/Phase5	MAP_PHASE5_APR09	4/30/2009	4/30/2009	4,037	126	4,456,448	a596908	cart	-rwxrwxrwx
09/campaigns/pick_a_pay/Phase5	PHASE5_MAIL_FILE_APR16_FINAL	5/13/2009	5/13/2009	3,634	17	958,464	a596908	cart	-rwxrwxrwx
09/campaigns/pick_a_pay/Phase5	PHASE5_MAIL_FILE_APR_16_W_SEI	5/29/2009	5/29/2009	4,070	128	4,718,592	a596908	cart	-rwxrwxrwx
Phase5	PHASE5_MAIL_FILE_DROP_APR17	4/29/2009	4/29/2009	41	2	16,384	a596908	cart	-rwxrwxrwx
Phase5	PHASE5_MAIL_FILE_DROP_MAY06	5/21/2009	5/21/2009	44	3	16,384	a596908	cart	-rwxrwxrwx
Phase5	PHASE5_MAIL_FILE_DROP_MAY13	5/13/2009	5/13/2009	263	18	98,304	a596908	cart	-rwxrwxrwx
Phase5	PHASE5_MAIL_FILE_DROP_MAY20	5/20/2009	5/20/2009	49	18	49,152	a596908	cart	-rwxrwxrwx
Phase5	PHASE5_MAIL_FILE_KEEP_APR17	4/17/2009	4/17/2009	3,321	17	884,736	a596908	cart	-rwxrwxrwx
Phase5	PHASE5_MAIL_FILE_KEEP_MAY06	5/21/2009	5/21/2009	3,557	17	933,888	a596908	cart	-rwxrwxrwx
Phase5	PHASE5_MAIL_FILE_KEEP_MAY12	5/13/2009	5/13/2009	3,557	18	958,464	a596908	cart	-rwxrwxrwx
Phase5	PHASE5_MAIL_FILE_KEEP_MAY13	5/20/2009	5/20/2009	3,294	18	884,736	a596908	cart	-rwxrwxrwx
Phase5	PHASE5_MAIL_FILE_KEEP_MAY20	5/28/2009	5/28/2009	3,245	20	933,888	a596908	cart	-rwxrwxrwx
Phase5	PHASE5_MAIL_WF_AUG14	8/14/2009	8/14/2009	3,681	72	2,949,120	a451481	cart	-rw-r--r--
Phase5	PHASE5_MAIL_WF_AUG14_FIN	8/14/2009	8/14/2009	3,681	85	3,473,408	a451481	cart	-rw-r--r--
Phase5	PHASE5_MAIL_WF_JUN03	6/5/2009	6/5/2009	3,245	64	2,408,448	a596908	cart	-rwxrwxrwx
Phase5	PHASE5_MAIL_WF_JUN30	7/6/2009	7/6/2009	3,245	75	2,883,584	a442587	cart	-rwxr--r--

rmat	label	varnum
	LTV	107
	LTV	107
	LTV	107
	LTV	107
	LTV	21
	LTV	108
	LTV	101
	LTV	108
	LTV	104
	LTV	104
	LTV	104
	LTV	104
	LTV	104
	LTV	98
	LTV	98
	LTV	98
	LTV	98
	LTV	110
	LTV	105
	LTV	105
	LTV	105
	LTV	105
	LTV	105
	LTV	104
	LTV	108
	LTV	104
	LTV	104
	LTV	104
	LTV	104
		22
		13
		13
		14
		8

The owner name can be looked up in /etc/passwd, and we can contact users about any data issues.

Dataset permissions may reveal possible issues
•Permissions (-rw-r--r--) do not alow group members cannot write to this dataset. Programs may fail when they are not run by user a451481.
•Other permissions (-rwxrwxrwx) allow everyone unlimited access to the datasets. Inexperienced users may accidently delete important information.

MOTIVATION FOR DATA PROFILING

- Initial motivation → Solve the immediate problem[1]
    1. Discover what SAS datasets are available.
    2. Describe the data structure and columns in the datasets.
    3. Show cases where data profiling meets specific needs.
    4. Provide a brief design for the generic data profiler program, which eventually became the Dataset_Profiler program.

- Current motivation → Create a data profiling solution
    1. The generic data profiler[2] was created to profile of any SAS dataset.
    2. Describe the data structure and columns, and show all category values for each reporting variable[3], plus statistics for each column.
    3. Instead of meeting specific needs for data profiling, show that these two programs[4] have all the features of a data profiler. Matching the definition of a data profiler would classify the programs as a good data profiling solution.
    4. Research was done to find a good definition of data profiling. A vendor-neutral definition[5] was got from TDWI (The Data Warehousing Institute).

[1] "Using Dictionary Tables to Explore SAS® Datasets"
[2] *Dataset_Profiler.sas*
[3] A reporting variable is defined as having its Number of Unique values < 300, and its Percent of Unique values <= 10%.
[4] *Dataset_Explorer.sas* and *Dataset_Profiler.sas*
[5] Philip Russom, "Raising the Bar for Data Profiling", What Works in Data Integration, Volume 29, pp. 2 – 5, (2010, TDWI).

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 metadata accurately describes the actual values in the source database.

Practice Area	Description	Enabler	Required Input
Data Profiling	Create a data inventory with profiles	Dataset_Profiler.sas	A dataset name
Data Discovery	Discover new and unknown data sources	Dataset_Explorer.sas	Directory names and/or Directory tree names
Data Monitoring	Re-profile and discover what has changed	Dataset_Explorer.sas Dataset_Profiler.sas	Two Explorer or Profile datasets, plus a program that compares two SAS datasets
Collaborative Profiling	Business people add meaning to the columns	Dataset_Explorer.sas Dataset_Profiler.sas	Collect comments about the values of the reporting variables

Figure 8. Four Practice Areas of Data Profiling

TEN BEST PRACTICES IN DATA PROFILING

Best Practices	Programs	Capabilities
1. Just do it! 2. Profile data thoroughly 3. Produce more through data profiles	Dataset_Profiler	• Easy to do; only requires a dataset location. • Efficient 2-pass algorithm is suitable for big data. • Statistics, owner, permissions, and actual values for reporting columns.
4. Discover and profile new data sources 5. Re-profile data as it evolves 6. Re-profile data daily via data monitoring	Dataset_Explorer Dataset_Profiler	• Searches directory trees to discover new data. • Compare Explorer datasets to find SAS timestamp changes, or data and structure changes. • Automate the comparisons. • Profile new datasets, and re-profile changed datasets.
7. Profile data across multiple IT systems 8. Collaborate through data profiles	Dataset_Profiler Dataset_Explorer	• Most databases are accessible by SAS/ACCESS • Profiler and Explorer programs can be adapted to databases by using the Libname access method and SAS In-Database technology. • “Seeing” the actual values of reporting variables encourages collaboration, since seeing is believing.
9. Map data as you discover and profile it	No	• Foreign keys are difficult to discover without rigid data naming conventions. SAS Web Report Studio maps foreign keys under the right conditions.
10. Support many practices with data profiling, discovery, and monitoring	Maybe	• Integration with some DI and DQ products may be possible as more databases and appliances support SAS In-database technology. • Integration with BI and DW is less likely to occur soon • SAS BI Platform integrates DI, DQ, BI, and DW.

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.
  - For details, see 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 for big 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 to determine whether a meaningful report could be produced from the dataset. In other words, did it have enough useful data to create a good analysis?

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 or blank 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

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 STATISTICS REPORT 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

PROGRAMMING DETAILS FOR DATASET EXPLORER

1. The program is well commented. Please read it for further details.
2. *%let*'s at the top of the program define the data locations and directories to search for SAS datasets.
3. Create a UNIX filename pipe to find all SAS datasets. Then read and process the results into a SAS dataset.
4. Create a libname for each directory that contains any SAS dataset.
5. Using the SAS Dictionary tables, get metadata from all datasets in your set of libnames.
6. Merge the UNIX information from step 3 with the SAS information from step 5.
7. Create the Excel reports.
8. Five SAS work datasets are downloaded to the PC, in case you want to do further analysis.

PROGRAMMING DETAILS FOR DATASET PROFILER

1. The program is well commented. Please read it for further details.
2. *%let*'s at the top of the program define the data locations and the SAS dataset to analyze.
3. Count missing, non-missing, and unique values for the dataset.
4. Transpose the dataset from step 3, which is a single row.
5. Process the transposed dataset to define all the variables for the "Column Report for Profiled Variables". Define reporting variables based upon number of unique values and % unique. Set Stats = "Y" for reporting variables.
6. Define macro variables to create the Proc Means analysis that will profile the dataset.
7. Run the Proc Means.
8. Process the Means output dataset. In simple terms, the large and sparse matrix of Means types is "squished" so that each value of a reporting variable is shown with column statistics for all numeric columns in the dataset.
9. Output the two Excel reports.
10. Filter the profile report to show what may be relevant data for your analysis. The text filters in Excel are very useful for searching and providing meaningful results.
11. The Proc Means output dataset is downloaded to SAS Work, in case you want to do further analysis.

PROGRAMS ARE AVAILABLE ONLINE

See the References section for the PowerPoint presentation with embedded SAS code.
Source code is available in the Credit Card section of the proceedings for IFSUG 2012:
http://www.ifsug.org/2012-proceedings

CONCLUSION

This paper has described a good data profiling solution, based upon the definition from TDWI. The visual elements showed what is possible with a data profiler based upon SAS data dictionary tables, and those results should speak for themselves. Program details were briefly described to assist those who decide to read the well commented code.

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-rofiling.aspx?sc_lang=en
- Phillip Julian, "Using Dictionary Tables to Explore SAS® Datasets", SESUG 2010 Proceedings, (2010), http://analytics.ncsu.edu/sesug/2010/PO23.Julian.pdf
- The presentation with source code is available from the SAS Community at http://www.sascommunity.org/wiki/File:Using_Dictionary_Tables_to_Profile_SAS_Datasets.pptx

CONTACT INFORMATION

I am interested in how this technology performs in the wild. Please feel free to share your experiences. Contact the author at:

Phillip Julian
Bank of America
Charlotte, NC
(469) 256-8463
julianp@acm.org
member.acm.org/~julianp