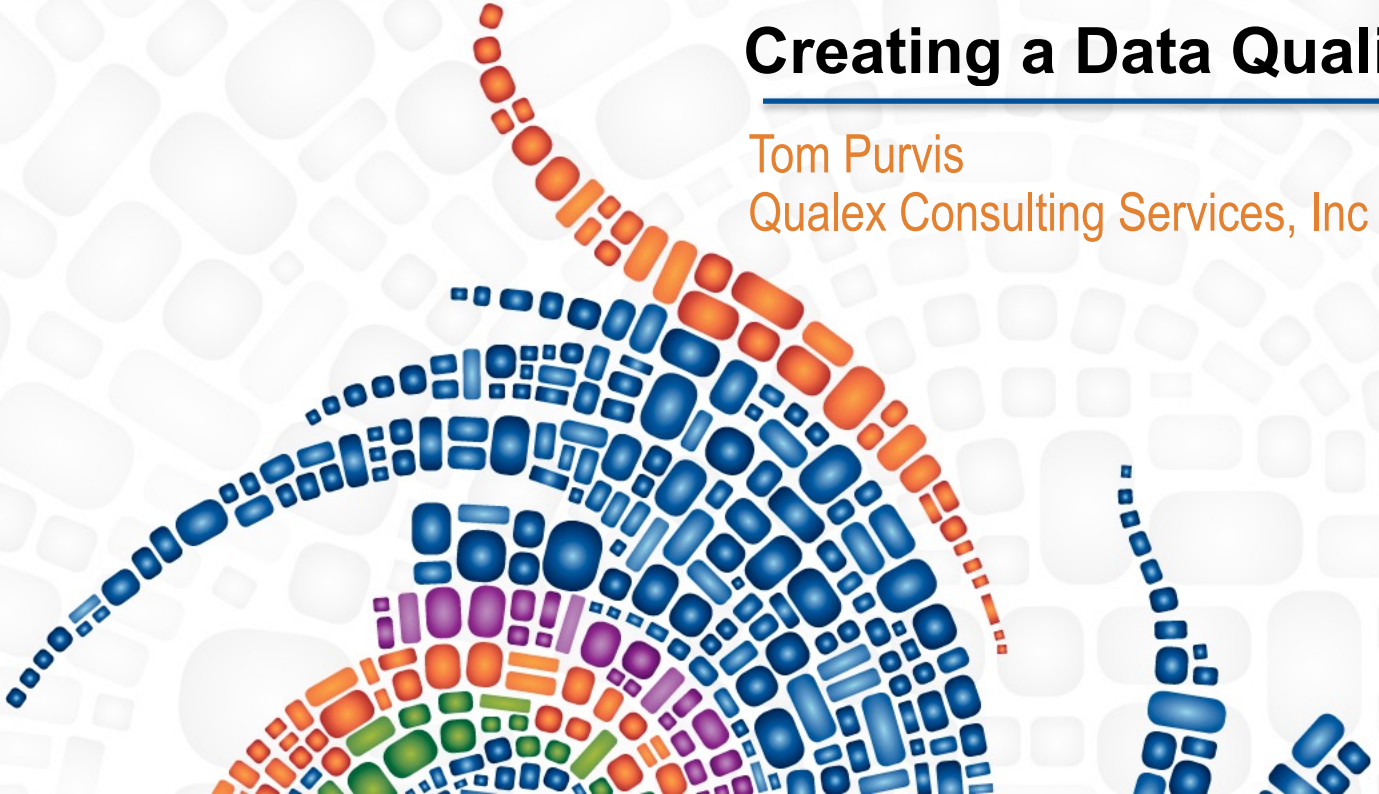
SAS® GLOBAL FORUM 2015

The Journey Is Yours

Creating a Data Quality Scorecard

Tom Purvis
Qualex Consulting Services, Inc



SAS GLOBAL FORUM

Session ID #3261-2015

How Do I Know If My Data Is Any Good?

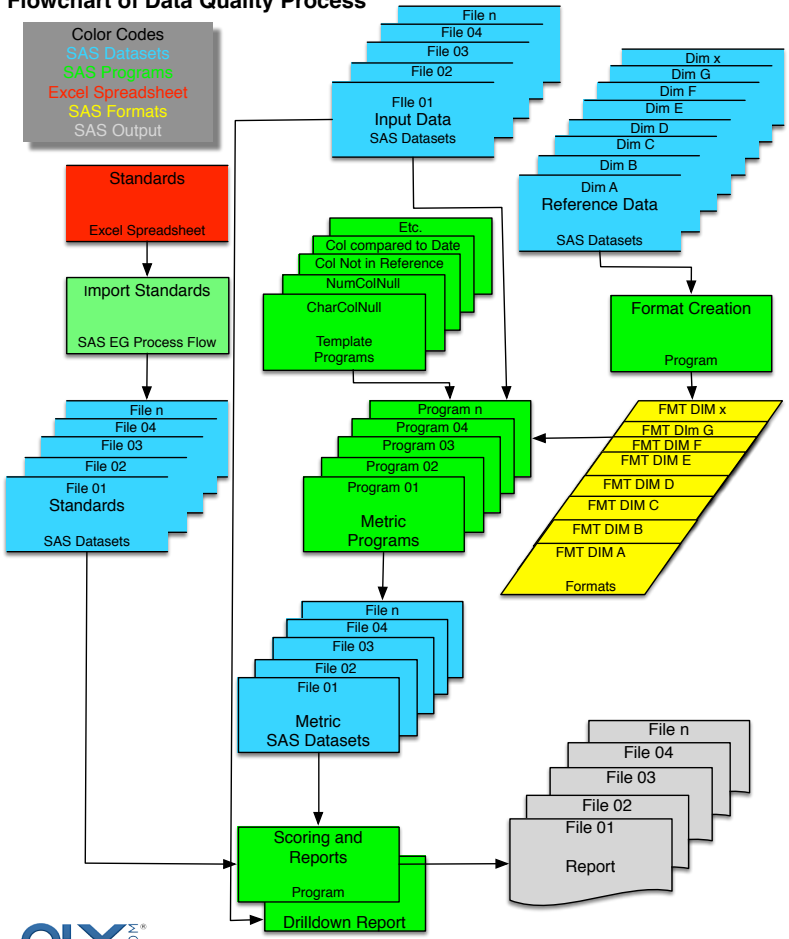
- Can I trust it?
- Is it consistent?
- Is one supplier of my data better or worse than others?
- What feedback can I give my data suppliers to help the process?
- Should this data have been loaded into the EDW?

The Fundamentals

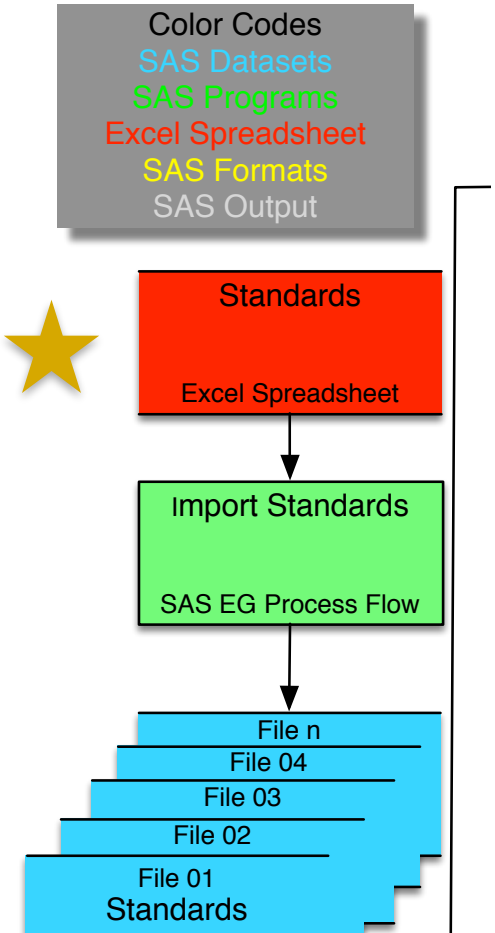
- Rules (green)
- Templates (green)
- Standards (red)
- Formats (yellow)
- Reports (gray)
- Trends
- Feedback

- Is customer number Blank?
- Is birthdate a proper date value greater than mm/dd/ccyy?
- Does the sum of Purchased Plus S&H Plus Tax = Total
- Is product code value a valid product code?
- Is the Sales Tax column computed correctly?

Flowchart of Data Quality Process

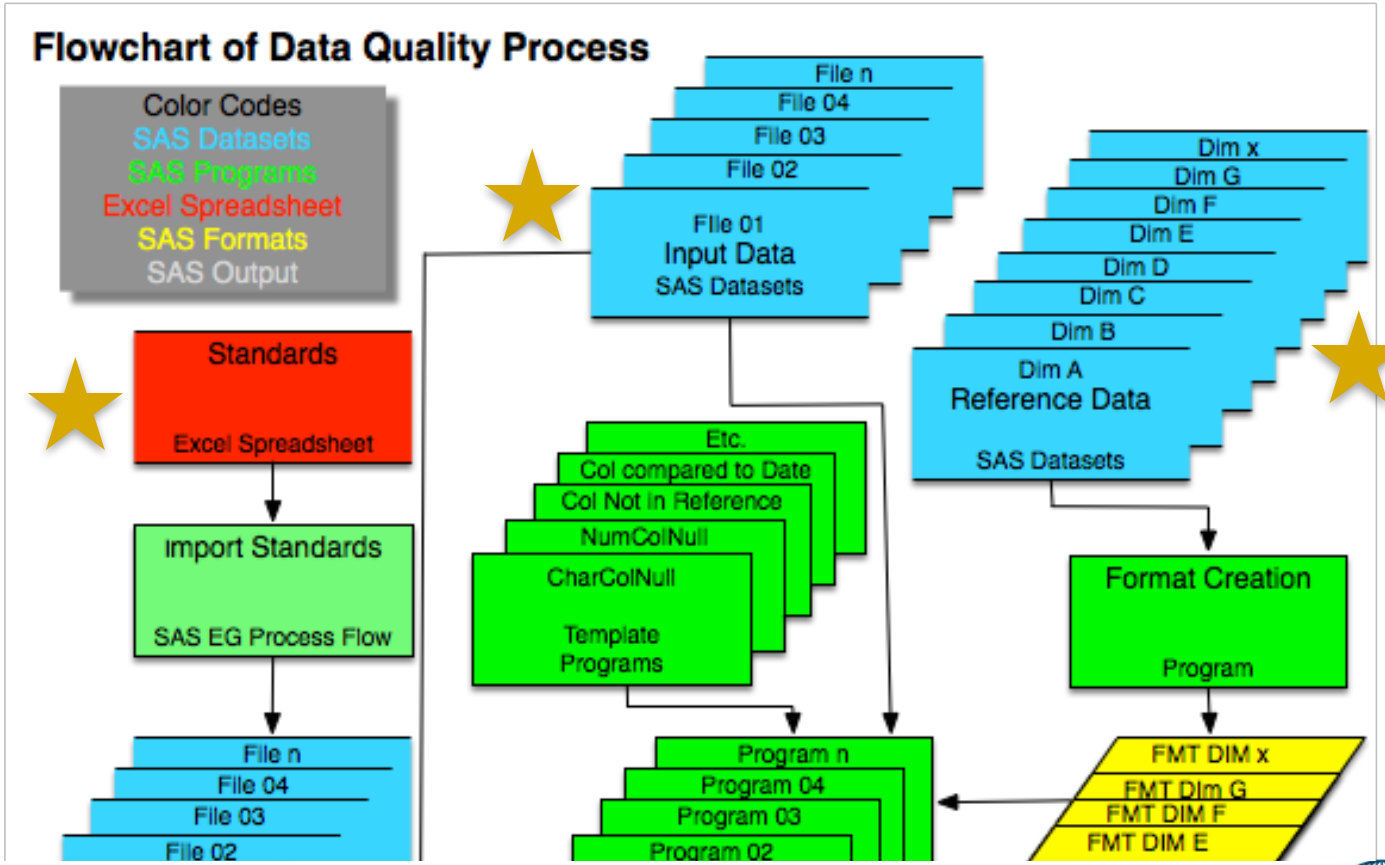


Flowchart of Data Quali



It Starts With A Spreadsheet

	A	C	D	E	F	G	H	I	J
1	Data_Field_Name	Include_In_Report	Metric_ID	Metric_Label	Green_Beg	Green_End	Blue_End	Red_end	Notes
2	Column_01	Yes	Rule 001	XXX ID is blank but XXX Name is not blank	0.00	10.00	20.00	100.00	
3	Column_01	Yes	Rule 002	Blank Values in Column	0.00	10.00	20.00	100.00	
4	Column_03	Yes	Rule 003	Blank Values in Column	0.00	10.00	20.00	100.00	
5	Column_07	Yes	Rule 004	Blank Values in Column	0.00	10.00	20.00	100.00	
6	Column_41	Yes	Rule 005	Blank Values in Column	0.00	10.00	20.00	100.00	
7	Column_42	Yes	Rule 006	Other XXX ID is blank but Other XXX Name is not blank	0.00	0.00	0.00	100.00	
8	Column_42	Yes	Rule 007	Blank Values in Column	0.00	10.00	20.00	100.00	
9	Column_44	Yes	Rule 008	Blank Values in Column	0.00	10.00	20.00	100.00	
10	Column_49	Yes	Rule 009	ZZZ_ID is blank but ZZZ_Name is not blank	0.00	0.00	0.00	100.00	
11	Column_49	Yes	Rule 010	Blank Values in Column	0.00	10.00	20.00	100.00	
12	Column_51	Yes	Rule 011	Blank Values in Column	0.00	10.00	20.00	100.00	
13	Column_55	Yes	Rule 012	Blank Values in Column	0.00	10.00	20.00	100.00	
14	Column_55	Yes	Rule 013	PTermination_Date GT Rundate	0.00	0.00	0.00	100.00	
15	Column_45	Yes	Rule 014	Effective_date GT Rundate	0.00	0.00	0.00	100.00	
16	Column_45	Yes	Rule 015	Blank Values in Column	0.00	10.00	20.00	100.00	
17	Column_46	Yes	Rule 016	XXXX_Effective_Date is Prior to Termination date	100.00	100.00	100.00	0.00	
18	Column_52	Yes	Rule 017	XXXX_ID values Blank but XXXX_Name is not Blank	100.00	100.00	100.00	0.00	
19	Column_52	Yes	Rule 018	Blank Values in Column	0.00	10.00	20.00	100.00	
20	Column_54	Yes	Rule 019	Blank Values in Column	0.00	10.00	20.00	100.00	
21	Column_57	Yes	Rule 020	XXXX_Type_Code values Blank but XXXX_Name is not Blank	100.00	100.00	0.00	0.00	


Flowchart of Data Quality Process
Color Codes
SAS Datasets
SAS Programs
Excel Spreadsheet
SAS Formats
SAS Output
File n
File 04
File 03
File 02
File 01
Input Data
SAS Datasets
Dim x
Dim G
Dim F
Dim E
Dim D
Dim C
Dim B
Dim A
Reference Data
SAS Datasets
Standards
Excel Spreadsheet
Import Standards
SAS EG Process Flow
Etc.
Col compared to Date
Col Not in Reference
NumColNull
CharColNull
Template
Programs
Format Creation
Program
File n
File 04
File 03
File 02
Program n
Program 04
Program 03
Program 02
FMT DIM x
FMT Dim G
FMT DIM F
FMT DIM E

Next We Create Some Templates

```
/* Column Date Compared to Date */
metric_id  = "&metric_id." ;
if &col. &comp. &date1. then
rowfailed = 1 ;
else rowfailed = 0 ;
output ;

%let metric_id = ;
%let col       = ;
%let comp      = ;
%let date1     = ;
```

```
/* Character Format */
metric_id  = "&metric_id." ;
if strip(&col.) NE " " and
   strip(put(&col.,&fmttest.)) = strip(&col.) then
rowfailed  = 1 ;
else
rowfailed  = 0 ;
output ;

%let metric_id = ;
%let col       = ;
%let fmttest   = ;
```

Making A Rule From A Template

```
%let metric_id = Bdate01;
%let col       =    birthdate;
%let comp      = LT ;
%let date1     = “01jan1900”d ;
%include “<path>/datecomp.sas” ;
```

```
%let metric_id = Transdate001 ;
%let col       =    transdate;
%let comp      = GT;
%let date1     = today()  ;
%include “<path>/datecomp.sas”  ;
```

Creating A Format

```
libname fmts  "/sasdata/data/validate/" ;

options fmtsearch=(fmts) ;
%let type    = C  ;
%let destin  = fmtZIP ;
%let source  = tbdim_zipcode ;

%let length  = 5 ;


data work.&source. ;
set  fmts.&source. ;
start = substr(strip(txtZIP9),1,&length.) ;
label = "Found" ;
run ;
```

```
proc sort data=work.&source. ;
by  start label   ;
run ;

data work.&source. ;
set  work.&source. ;
by  start label   ;
if last.start ;
run ;


data fmts.&destin. ;
set  work.&source.  ;

fmtname = "&destin." ;
type = "&type." ;

end   = start ;
length = &length. ;
```

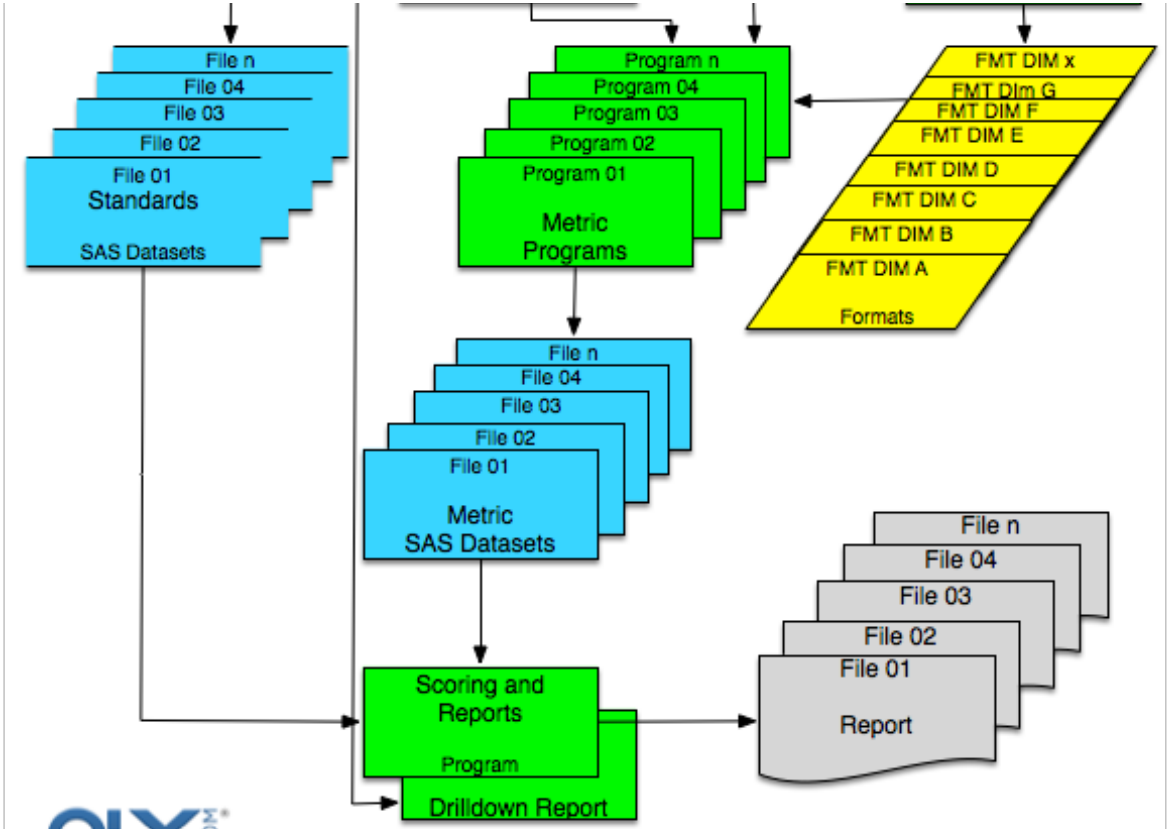
Creating A Format con't

```
/*if length(strip(&start.)) NE &length.  then delete ;*/

keep start end label type length fmtname ;

run ;

proc format cntlin=fmts.&destin. library=fmts.formats  ;
run ;
```


File n
File 04
File 03
File 02
File 01
Standards
SAS Datasets
Program n
Program 04
Program 03
Program 02
Program 01
Metric
Programs
FMT DIM x
FMT DIM G
FMT DIM F
FMT DIM E
FMT DIM D
FMT DIM C
FMT DIM B
FMT DIM A
Formats
File n
File 04
File 03
File 02
File 01
Metric
SAS Datasets
File n
File 04
File 03
File 02
File 01
Report
Scoring and
Reports
Program
Drilldown Report

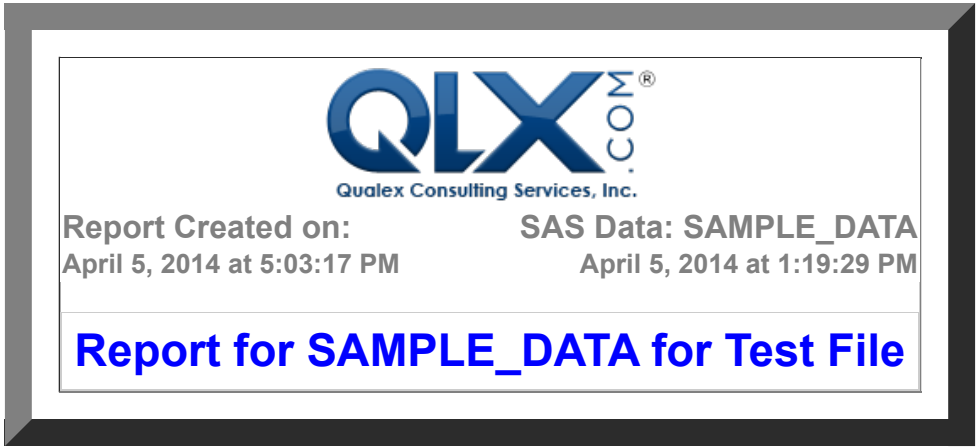
Creating A Drill Through Report

```
machine = "http://111.222.333.444:8080/" ;
engine  = "SASStoredProcess/do?" ;
progpath = "_program=SBIP://METASERVER%2F<folder>%2F<folder>%2F<folder>%2F" ;
program  = "ReviewOfInputData" ;
paramA   = "colname="  ;
odsstyle = "&_odsstyle." ;
 ;
```

Creating A Drill Through Report con't

```
htmlvar = "<a href=" ||
""" || /* beginning single quote for href */
trim(machine) ||
trim(engine) ||
trim(progpath) ||
trim(program) || '&' ||
trim(paramA)  ||  strip(Data_field_name)   ||
'&_odsstyle='       || strip(odsstyle) ||
"'>" ;

Metric_ID_linked = htmlvar ||  strip(Metric_ID) || '</a>' ;
```



File Header Information

	SAS Dataset Name
	SAMPLE_DATA
Record Length	2,248
Number of Variables	42
Number of Records	9,782
Date Created	05APR2014:13:19:29

Summary Statistics Report

Score	Number of Columns	Number of Metrics
Pass	21	26
Check	1	1
Fail	11	11
	33	38

Column Summary Report

Field Name	Score	Number of Metrics
Column_01	Pass	2
Column_03	Pass	1
Column_04	Fail	1
Column_06	Pass	1
Column_07	Pass	1
Column_08	Pass	1
Column_09	Pass	2
Column_09	Check	1
Column_14	Fail	1
Column_15	Pass	1
Column_16	Pass	2
Column_18	Pass	1
Column_20	Pass	1
Column_25	Fail	1
Column_28	Fail	1
Column_29	Fail	1
Column_32	Fail	1
Column_33	Fail	1
Column_36	Pass	1
Column_37	Pass	1
Column_38	Fail	1
Column_41	Pass	1
Column_42	Pass	2
Column_44	Pass	1
Column_45	Pass	1
Column_48	Fail	1
Column_49	Pass	2
Column_51	Fail	1
Column_52	Pass	1
Column_54	Pass	1
Column_55	Pass	1

Column and Metric Details Report

Column Name	Column Name	Metric Description	Rows Failed	Rows Processed	Percent Failed	Score
RULEXX	Column_01	Vendor ID is blank but Vendor Name is not blank	0	9,782	0.0%	Pass
RULEXX	Column_01	Blank Values in Column	111	9,782	1.1%	Pass
RULEXX	Column_03	Blank Values in Column	111	9,782	1.1%	Pass
RULEXX	Column_07	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_41	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_42	Other Vendor ID is blank but Other Vendor Name is not blank	0	9,782	0.0%	Pass
RULEXX	Column_42	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_44	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_49	Product_Package_ID is blank but Product_Package_Name is not blank	0	9,782	0.0%	Pass
RULEXX	Column_49	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_51	Blank Values in Column	9,782	9,782	100%	Fail
RULEXX	Column_55	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_55	Customer_Termination_Date GT Rundate	9,634	9,782	98%	Fail
RULEXX	Column_45	Blank Values in Column	1	9,782	0.0%	Pass
RULEXX	Column_52	Blank Values in Column	0	9,782	0.0%	Pass
RULEXX	Column_54	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_57	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_48	Blank Values in Column	9,782	9,782	100%	Fail
RULEXX	Column_08	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_36	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_37	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_38	Blank Values in Column	9,782	9,782	100%	Fail
RULEXX	Column_14	Blank Values in Column	9,782	9,782	100%	Fail
RULEXX	Column_32	Blank Values in Column	9,782	9,782	100%	Fail
RULEXX	Column_04	Blank Values in Column	9,782	9,782	100%	Fail
RULEXX	Column_06	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_09	Customer_Date_of_Birth >= Rundate	16	9,782	0.2%	Pass
RULEXX	Column_09	Customer_Date_of_Birth < 01Jan1900	148	9,782	1.5%	Check
RULEXX	Column_09	Blank Values in Column	148	9,782	1.5%	Pass
RULEXX	Column_15	Blank Values in Column	0	9,782	0.0%	Pass

Top 5 Frequency for Values of Customer_ID

Obs	Customer_ID	COUNT	PERCENT
1	30698	11	0.11245
2	51603	11	0.11245
3	81612	11	0.11245
4	14369	10	0.10223
5	45532	10	0.10223

Top 5 Frequency for Length of Customer_ID

Obs	Customer_ID_len	COUNT	PERCENT
1	5	9782	100

Top 5 Frequency for Values of Vendor_ID

Obs	Vendor_ID	COUNT	PERCENT
1		111	1.1347
2	AAAAAA	9671	98.8653

Top 5 Frequency for Length of Vendor_ID

Obs	Vendor_ID_len	COUNT	PERCENT
1	6	9671	98.8653
2	1	111	1.1347

Top 5 Frequency for Values of Customer_gender

Obs	Customer_Gender	COUNT	PERCENT
1	F	4891	50
2	M	4891	50

Top 5 Frequency for Length of Customer_gender

Obs	Customer_gender_len	COUNT	PERCENT
1	1	9782	100

Top 5 Frequency for Values of Customer_Date_of_Birth

Obs	Customer_Date_of_Birth	COUNT	PERCENT
1	.	148	1.5130
2	06/01/1936	989	10.1104
3	11/20/1953	8	0.0818
4	06/28/1954	7	0.0716
5	10/12/1953	7	0.0716

Top 5 Frequency for Length of Customer_Date_of_Birth

Obs	Customer_Date_of_Birth_len	COUNT	PERCENT
1	12	9782	100

How to Set the Background Colors for the Score

```
proc format ;
 value score
   1 = "Pass"
   2 = "Check"
   3 = "Fail"
   ;
value colorsc
  1 = 'Light Green'
   2 = 'Cyan'
   3 = 'Very Light Red'
   ;
run ;
```

```
proc print data=work.filesummary label noobs ;
var score / style={background=colorsc.}  ;

var Columns  metrics ;
sum columns metrics ;
format score score. ;
label Score = "Score"
      Columns = "Number of Columns"
                Metrics = "Number of Metrics"
                ;
format columns
       metrics   comma20.0
                 ;
run ;
```

Tips, Tricks, and Traps

- Using %Includes to call templates
- Using Formats to do lookups
- Using Formats to set colors
- Creating columns linked to stored processes
- RAM Usages
- Naming Conventions

Trending the Results

- If you make the SAS metric datasets permanent what could you do?
  - Trend each Rule over time.
  - Use the trends to revise standards
  - Use trends as feedback to the data supplier.
  - Use drill down and trends to improve data collection process
  - Compare one supplier to another

SAS® GLOBAL FORUM

April 26-29
Dallas, TX