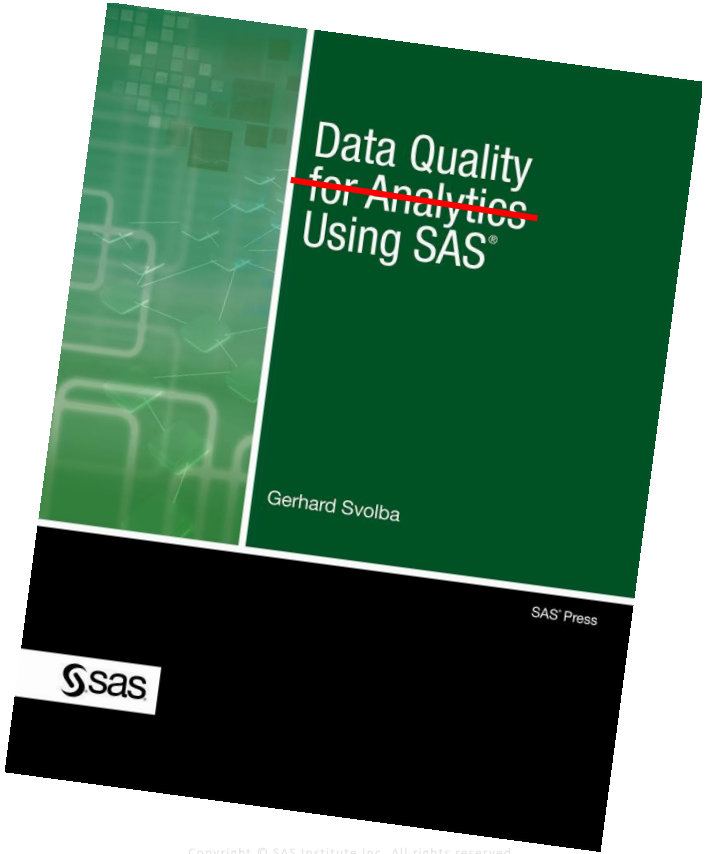
Datenqualität leicht gemacht mit SAS Data Quality

SAS Club 2018, Wien

Datenqualität – längst erledigt?

sas

Datenqualität – längst erledigt?



Funktionsübersicht - Datenqualität

Data Profiling
Datenprobleme verstehen

Standardization
Sicherstellung der Datenkonsistenz

Data Monitoring
Überwachen von Geschäftsregeln

Entity Resolution
Erstellung zusammengefaßter Entitäten

Business Data Glossary
Verwalten von Geschäftsbezeichnungen und Glossar

Master Data
Stammdaten-Management

Reference Data Management
Referenzdatenpflege durch Fachbereich

Enrichment
Anreichern von Daten

Data Governance & Stewardship
Zusammenarbeit Fachbereich & IT

Regressionsrechnung
(linenar, log., polynomial, etc.)

Mittelwert
Abweichung
MIN/MAX
MODE

Weitere statistische Methoden

Profiling: Analyse der Daten

Metadata Validierung

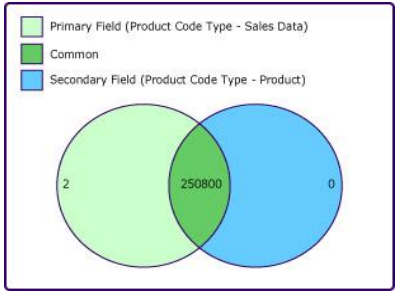
Prüfung der Datensätze nach verschiedenen Kriterien
(z.B. Anzahl unterschiedlicher Einträge, Primärschlüsselkandidat, etc.)

Pattern Analyse

Ermittelt Muster (Pattern) in Feldern, Nummern werden als 9 und Buchstaben als A dargestellt. Typische Pattern sind die Kreditkartennummer, Sozialversicherungsnummer, etc.)

Statistiken

Ermittelt Werte wie Min., Max. Mittelwert, Standardabweichung, etc.



Häufigkeitszählung

Listet die Anzahl von Einträgen auf.

Regelprüfung

Überprüfung der Einhaltung von spezifischen Geschäftsregeln.

Beziehungsanalyse

Zeigt an, ob Primär- & Fremdschlüssel Relationen konsistent sind oder wie hoch die redundante Datenhaltung ist. Hiermit können auch Ausreißer entdeckt werden, wenn z.B. Redundanz erwartet wird.

METRIC NAME	METRIC VALUE
Data Type	double
Primary Key Candidate	no
Unique Count	1140
Uniqueness	70.11
Pattern Count	(not applicable)
Minimum Value	-223000
Maximum Value	9999999
Minimum Length	(not applicable)
Maximum Length	(not applicable)
Null Count	2
Blank Count	(not applicable)
Actual Type	double
Count	1628
Data Length	53 bit
Mean	114348.170972
Median	4888499.5
Mode	0
Non-Null Count	1626
Nullable	YES
Ordinal Position	7
Decimal Places	0
Standard Deviation	429438.361236
Standard Error	10649.778281

Match Codes – PROC DQMATCH

Zusammenhänge erkennen

Basic Settings | Advanced Properties | Preview | Node Connections | Log

Daten clustern

View Value...

	ID	Name	Strasse	Name_MatchCode	Cluster_ID	
8	8	Herr Prof. Josef Steigenmann von Hüls	Linden 12	2#W4$$$$$$$7@4_G$$$$	1	System 1 – z.B. Marketing
9	9	Josef Steigenmann v. Hülls	Unter den Linden 12	2#W4$$$$$$$7@4_G$$$$	1	System 2 – z.B. ERP
10	10	Josef Steigenmann v. Hüls	Unter den Linden 12a	2#W4$$$$$$$7@4_G$$$$	1	System 3 – z.B. WEB
11	11	Joseph Steigenmann von Hüls	Unter den Linden 12	2#W4$$$$$$$7@4_G$$$$	1	System 4 – z.B. Holding
12	12	Prof. Dr. med. Josef Steigenmann von Hüls	Unter den Linden 12	2#W4$$$$$$$7@4_G$$$$	1	System 5 – z.B. FI
13	13	Sepp Steigenmann v. Hüls	Den Linden 12	2#W4$$$$$$$7@4_G$$$$	1	System 6 – z.B. Partner
14	14	Dipl-Inf Maximilian Mustermann	o 3 12	B#4~_YB&P$$B&X$$$$$$	2	
15	15	Dipl.-Inf. Max Mustermann	O 3. 12	B#4~_YB&P$$B&X$$$$$$	2	
16	16	Diplom-Informatiker Max Mustermann	O 3, 12	B#4~_YB&P$$B&X$$$$$$	2	
17	17	Herr Dipl-Inf Max Mustermann	o 3, 12	B#4~_YB&P$$B&X$$$$$$	2	
18	18	Mustermann, Max	o 3, 12	B#4~_YB&P$$B&X$$$$$$	2	
19	19	Hans Meier	Konrad Adenauerstesse 42	B_7&$$$$$$$2&P4$$$$$	3	
20	20	Hans Maier	K-Adenauer Str. 42	B_7&$$$$$$$2&P4$$$$$	3	
21	21	Herr Dr. Hans Meier	Konrad-Adenauer Str 42	B_7&$$$$$$$2&P4$$$$$	3	

Match Codes - Stammdaten

Zusammenhänge erkennen

Program* | Log | Output Data (2) | Results

Save | Run | Stop | Selected Server: SASApp (Connected) | Analyze Program | Expo

```
/* Define QKB locale */
option DQLOCALE=(enusa);

/* Create the input data set. */
data cust_db;
   length customer $ 22;
   length address $ 31;
   input customer $char22. address $char31.;
datalines;
Bob Beckett            392 S. Main St. PO Box 2270
Robert E. Beckett      392 S. Main St. PO Box 2270
Rob Beckett            392 S. Main St. PO Box 2270
Paul Becker            392 N. Main St. PO Box 7720
Bobby Becket           392 Main St.
Mr. Robert J. Beckeit  P. O. Box 2270 392 S. Main St.
Mr. Robert E Beckett   392 South Main Street #2270
Mr. Raul Becker        392 North Main St.
;
run;
```

Match Codes - Stammdaten

Zusammenhänge erkennen

Program* | Log | Output Data (2) | Results

Save | Run | Stop | Selected Server: SASApp (Connected) | Analyze Program | Expo

```
/* Define QKB locale */
option DQLOCALE=(enusa);

/* Create the input data set. */
data cust_db;
   length customer $ 22;
   length address $ 31;
   input customer $char22. address $char31.;
datalines;
Bob Beckett            392 S. Main St. PO Box 2270
Robert E. Beckett      392 S. Main St. PO Box 2270
Rob Beckett            392 S. Main St. PO Box 2270
Paul Becker            392 N. Main St. PO Box 7720
Bobby Becket           392 Main St.
Mr. Robert J. Beckeit  P. O. Box 2270 392 S. Main St.
Mr. Robert E Beckett   392 South Main Street #2270
Mr. Raul Becker        392 North Main St.
;
run;

/* Run the DQMATCH procedure. */
proc dqmatch data=cust_db out=out_db1 matchcode=match_cd
   cluster=clustergrp locale='ENUSA';
   criteria matchdef='Name' var=customer;
   criteria matchdef='Address' var=address;
run;
```

Match Codes - Stammdaten

Zusammenhänge erkennen

Program* | Log | Output Data (2) | Results

Save | Run | Stop | Selected Server: SASApp (Connected) | Analyze Program | Expo

```
/* Define QKB locale */
option DQLOCALE=(enusa);

/* Create the input data set. */
data cust_db;
   length customer $ 22;
   length address $ 31;
   input customer $char22. address $char31.;
datalines;
Bob Beckett              392 S. Main St. PO Box 2270
Robert E. Beckett        392 S. Main St. PO Box 2270
Rob Beckett              392 S. Main St. PO Box 2270
Paul Becker              392 N. Main St. PO Box 7720
Bobby Becket             392 Main St.
Mr. Robert J. Beckeit    P. O. Box 2270 392 S. Main St.
Mr. Robert E Beckett     392 South Main Street #2270
Mr. Raul Becker          392 North Main St.
;
run;

/* Run the DQMATCH procedure. */
proc dqmatch data=cust_db out=out_db1 matchcode=match_cd
   cluster=clustergrp locale='ENUSA';
   criteria matchdef='Name' var=customer;
   criteria matchdef='Address' var=address;
run;
```

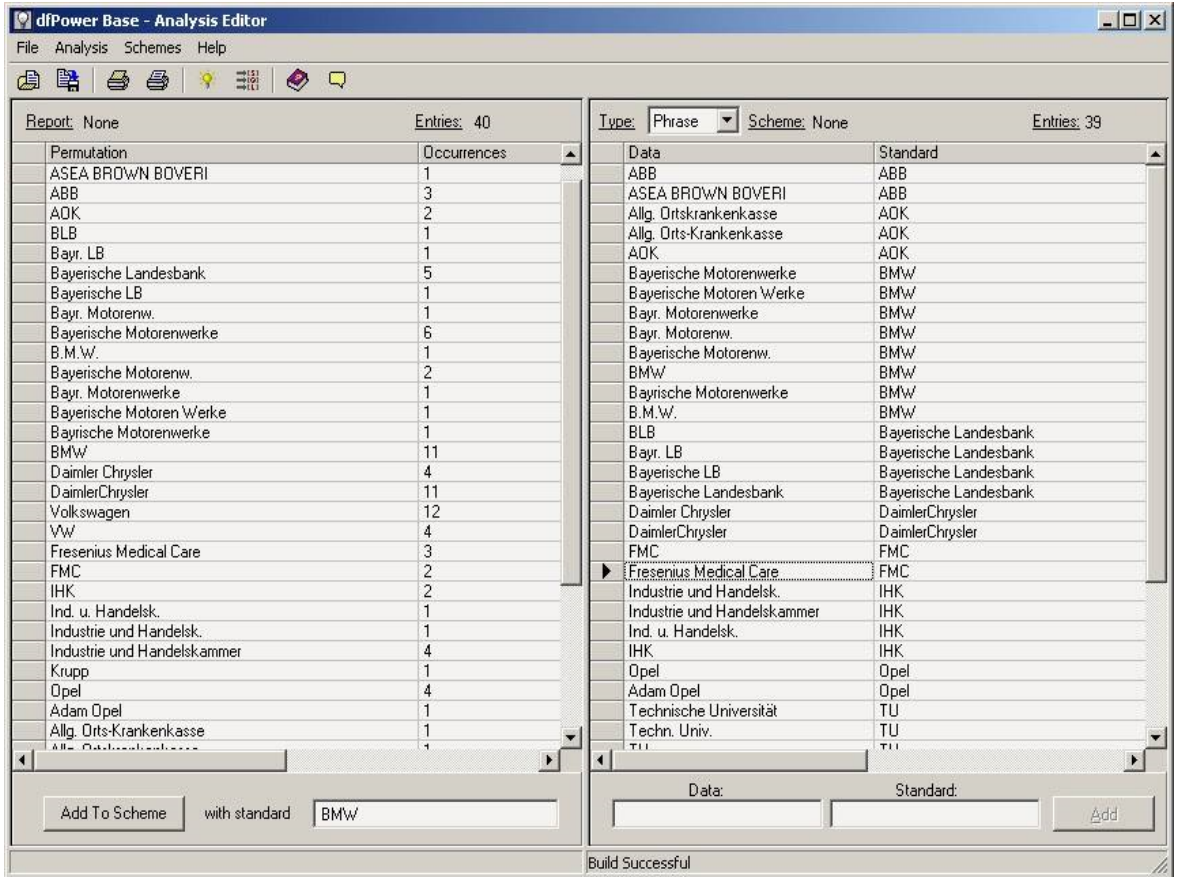
OUT_DB1

Filter and Sort | Query Builder | Where | Data | Describe | Graph | Analyze | Export | Send To

	customer	address	MATCH_CD	CLUSTERG...
1	Mr. Robert J. Beckeit	P. O. Box 2270 392 S. Main St	M3~$$$$M@M$$$$$!K-H$$BP$$HHI0$$	1
2	Bob Beckett	392 S. Main St. PO Box 2270	M3~$$$$M@M$$$$$!K-H$$BP$$HHI0$$	1
3	Rob Beckett	392 S. Main St. PO Box 2270	M3~$$$$M@M$$$$$!K-H$$BP$$HHI0$$	1
4	Mr. Robert E Beckett	392 South Main Street #2270	M3~$$$$M@M$$$$$!K-H$$BP$$HHI0$$	1
5	Robert E. Beckett	392 S. Main St. PO Box 2270	M3~$$$$M@M$$$$$!K-H$$BP$$HHI0$$	1
6	Paul Becker	392 N. Main St. PO Box 7720	M3Y$$$$NW$$$$$$!K-H$$BP$$IIH0$$	.
7	Bobby Becket	392 Main St.	M3~$$$$M@M$$$$$!K-H$$BP$$$$$$$$	.
8	Mr. Raul Becker	392 North Main St.	M3Y$$$$YW$$$$$$!K-H$$BP$$$$$$$$	.

Standardisierung – PROC DQSCHEME

Zusammenhänge erkennen



Standardisierung – PROC DQSCHEME

Zusammenhänge erkennen

```
/* Create the input data set. */
data vendors;
   input city $char16. state $char22. company $char34.;
datalines;
Detroit          MI                     Ford Motor
Dallas           Texas                  Wal-mart Inc.
Washington       District of Columbia   Federal Reserve Bank
SanJose          CA                     Wal mart
New York         New York               Ernst & Young
Virginia Bch     VA                     TRW INC - Space Defense
Dallas           TX                     Walmart Corp.
San Francisco    California             The Jackson Data Corp.
New York         NY                     Ernst & Young
Washington       DC                     Federal Reserve Bank 12th District
New York         N.Y.                   Ernst & Young
San Francisco    CA                     Jackson Data Corporation
Atlanta          GA                     Farmers Insurance Group
RTP              NC                     Kaiser Permanente
New York         NY                     Ernest and Young
Virginia Beach   VIRGINIA               TRW Space & Defense
Detroit          Michigan               Ford Motor Company
San Jose         CA                     Jackson Data Corp
Washington       District of Columbia   Federal Reserve Bank
Atlanta          GEORGIA                Target
;
run;
```

Standardisierung – PROC DQSCHEME

Zusammenhänge erkennen

```
/* Create the input data set. */
data vendors;
   input city $char16. state $char22. company $char34.;
datalines;
Detroit          MI                    Ford Motor
Dallas           Texas                 Wal-mart Inc.
Washington       District of Columbia  Federal Reserve Bank
SanJose          CA                    Wal mart
New York         New York              Ernst & Young
Virginia Bch     VA                    TRW INC - Space Defense
Dallas           TX                    Walmart Corp.
San Francisco    California            The Jackson Data Corp.
New York         NY                    Ernst & Young
Washington       DC                    Federal Reserve Bank 12th District
New York         N.Y.                  Ernst & Young
San Francisco    CA                    Jackson Data Corporation
Atlanta          GA                    Farmers Insurance Group
RTP              NC                    Kaiser Permanente
New York         NY                    Ernest and Young
Virginia Beach   VIRGINIA              TRW Space & Defense
Detroit          Michigan              Ford Motor Company
San Jose         CA                    Jackson Data Corp
Washington       District of Columbia  Federal Reserve Bank
Atlanta          GEORGIA               Target
;
run;
```

```
 /* Create the analysis data set. */
proc dqscheme data=vendors;
   create analysis=a_state
          matchdef='State (Scheme Build)'
          var=state
          locale='ENUSA';
run;
```

Standardisierung – PROC DQSCHEME

Zusammenhänge erkennen

```
/* Create the input data set. */
data vendors;
   input city $char16. state $char22. company $char34.;
datalines;
Detroit          MI                     Ford Motor
Dallas           Texas                  Wal-mart Inc.
Washington       District of Columbia   Federal Reserve Bank
SanJose          CA                     Wal mart
New York         New York               Ernst & Young
Virginia Bch     VA                     TRW INC - Space Defense
Dallas           TX                     Walmart Corp.
San Francisco    California             The Jackson Data Corp.
New York         NY                     Ernst & Young
Washington       DC                     Federal Reserve Bank 12th District
New York         N.Y.                   Ernst & Young
San Francisco    CA                     Jackson Data Corporation
Atlanta          GA                     Farmers Insurance Group
RTP              NC                     Kaiser Permanente
New York         NY                     Ernest and Young
Virginia Beach   VIRGINIA               TRW Space & Defense
Detroit          Michigan               Ford Motor Company
San Jose         CA                     Jackson Data Corp
Washington       District of Columbia   Federal Reserve Bank
Atlanta          GEORGIA                Target
;
run;
```

```
/* Create the analysis data set. */
proc dqscheme data=vendors;
   create analysis=a_state
          matchdef='State (Scheme Build)'
          var=state
          locale='ENUSA';
run;
```

A_STATE

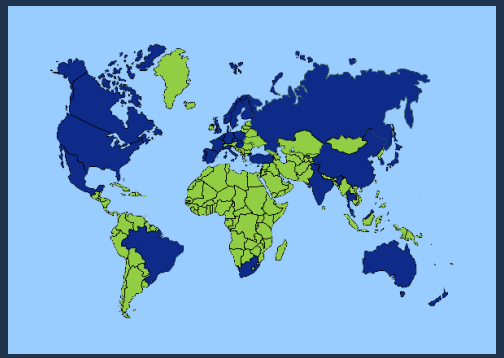
Filter and Sort | Query Builder | Where | Data

	COUNT	state	CLUSTER
1	2	District of Columbia	1
2	1	DC	1
3	1	MI	2
4	1	Michigan	2
5	1	GA	3
6	1	GEORGIA	3
7	3	CA	4
8	1	California	4
9	1	NC	.
10	2	NY	5
11	1	N.Y.	5
12	1	New York	5
13	1	VA	6
14	1	VIRGINIA	6
15	1	TX	7
16	1	Texas	7

Quality Knowledge Base

"behind the magic"

SAS Data Quality



SAS Quality Knowledge Base (QKB)

- **Regeln, Vokabeln und Prozesse** für Datenqualität, Datenintegration und Masterdatenmanagement:
  - Standardisierung (Parsing)
  - Mustervalidierung (Matching)
  - Geschlechtsanalyse (Identifizierung)
  - Extraktion
- Normalisierung von **Adressstandards** in 30+ Sprachen und Ländern:
  - Name (Organisation)
  - Merkmale (z.B. E-Mail)
  - Adresse (Lokalität)
- Optionen für **Datenanreicherung**:
  - Ergänzung von fehlenden Informationen
  - Anbindung von Web Services (Geocoding, D&B, Google Maps …)
- Datenqualitätsoptionen sind **flexibel und ausbaufähig**