Rechtzeitig mit SAS ein Bild über die Qualität der Analysedaten erhalten

Gerhard Svolba
SAS Austria
Mariahilfer Straße 116
A-1070 Wien
gerhard.svolba@sas.com

Zusammenfassung

Klarerweise will man über die Qualität seiner Analysedaten Bescheid wissen, BEVOR man viel Zeitaufwand in die Analyse gesteckt hat. Die analytischen Werkzeuge von SAS und JMP helfen, die Qualität Ihrer Analysedaten darzustellen und zu verbessern.

In diesem Beitrag finden Sie praktische Beispiele, wie SAS Base, SAS/STAT und SAS/ETS nicht nur für die Datenanalyse selbst verwendet werden kann, sondern diese Werkzeuge Aussagen über den Zustand Ihrer Daten liefern und Ihnen helfen die Datenqualität zu verbessern.

SAS/STAT erlaubt es, systematische Muster in Ihren Daten aufzudecken, z. B. bei fehlenden Werten oder Fehlern. Analytische Methoden in SAS/STAT und SAS/ETS ermöglichen, individuelle Validierungslimits zu definieren und Ausreißer in Zeitreihendaten zu erkennen. Die Prozeduren TIMESERIES und EXPAND erlauben fehlende Werte in Zeitreihendaten zu ersetzen und die Kontinuität einer Zeitreihe sicherzustellen. JMP bietet viele Möglichkeiten des interaktiven graphischen Profilings, um komplexe Ausreißer-Strukturen zu entdecken.

Dieser Vortrag basiert auf dem SAS-Press Buch „Data Quality for Analytics Using SAS" von Svolba [1]. Weitere Details und Screenhots finden Sie unter http://www.sascommunity.org/wiki/Data_Quality_for_Analytics.

Schlüsselwörter: Datenqualität, Datenaufbereitung, fehlende Werte, Missing Values, Validierung, JMP, TIMESERIES-Prozedur, EXPAND-Prozedur

1 Idee und Hintergrund von „Data Quality for Analytics"

„Data Quality for Analytics" und „Data Preparation for Analytics" definieren sich als "Das gesamte Ökosystem (Entscheidungen, Kriterien, Datenaufbereitungsschritte, fachliche Überlegungen) das zwischen der FACHLICHEN FRAGESTELLUNG und dem FINALEN ANALYTISCHEN MART liegt. Abbildung 1 illustriert diesen Zusammenhang.

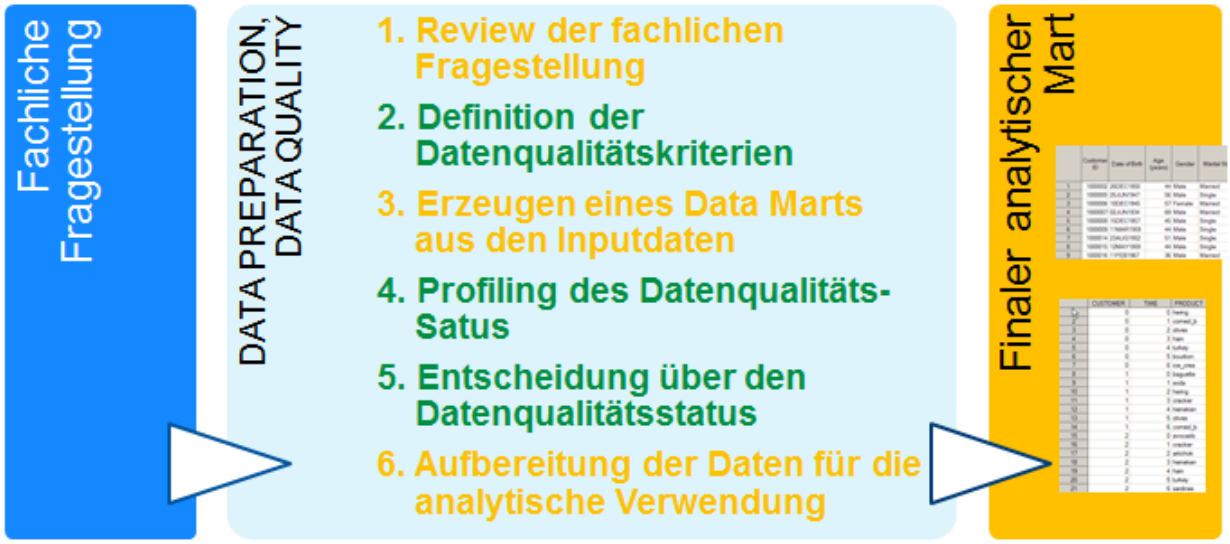

Abbildung 1: Einbettung von Data Quality und Data Preparation for Analytics

Die Punkte 1.) Review der fachlichen Fragestellung, 3.) Erzeugen eines Data Marts aus den Inputdaten und 6). Aufbereitung der Daten für die analytische Verwendung sind den Bereich „**Data Preparation for Analytics“** zuzuordnen. Siehe dazu auch [2] und [3].

Die Punkte 2.) Definition der Datenqualitätskriterien, 4.) Profiling des Datenqualitäts-Status und 5.) Entscheidung über den Datenqualitätsstatus fallen in den Bereich **Data Quality for Analytics.** Für mehr Details siehe [1].

Dieser Beitrag fokussiert sich auf den Punkt „Profiling des Datenqualitätsstatus“. Zuvor werden kurz die wichtigsten Datenqualitätskriterien skizziert.

2 Kriterien für Datenqualität aus analytischer Sicht

Aus analytischer Sicht sind folgende Kriterien für Datenqualität wichtig:

2.1 Datenverfügbarkeit

2.1.1 Historische Snapshots

Aus analytischer Sicht geht es nicht nur um aktuelle Daten sondern auch meist um historische Daten oder genauer um „Snapshots“ der Daten zu Zeitpunkten in der Vergangenheit. Die Bereitstellung dieser „historischen Sichtweisen“ ist in vielen Fällen problematisch, da operative Systeme aber auch manche Data Warehouse Systeme diese nicht speichern.

2.1.2 Laufende Verfügbarkeit

Ein weiterer Punkt ist die Gewährleistung der fortlaufenden Verfügbarkeit der Daten. Werden Analysen regelmäßig wiederholt oder Logiken als Ergebnis der Analyse regelmäßig angewandt, so ist bei der Datenauswahl auf die Möglichkeit zu achten, dass die Analysedaten auch in Zukunft in der gleichen Qualität und gleichen Definition bereitgestellt werden können.

2.1.3 Richtiger Granularitätslevel

Ein weiterer wichtiger Punkt ist der Granularitätslevel der Daten. Oft sind Daten zwar verfügbar, allerdings nur in aggregierter Form. Für die meisten statistischen Analysen sind die Daten aber als Einzeldaten pro Analysesubjekt nötig.

2.2 Datenmenge

Für aussagekräftige Analysen ist ein Mindestmaß an Datenmenge nötig. Dies betrifft in den meisten Fällen die Anzahl der Analyse-Subjekte, die Anzahl der Ereignisse und die Länge des Beobachtungszeitraumes. Methoden der Stichprobenplanung ermöglichen hier eine Aussage der minimal benötigten Datenmenge.

2.3 Datenvollständigkeit

Die Vollständigkeit der Daten bzw. die Anzahl der fehlenden Werte ist ein wichtiges Gütekriterium für die Datenverwendbarkeit. Hier ist insbesondere die Unterscheidung in zufällige oder systematisch fehlende Werte und das Aufdecken von Mustern in den fehlenden Werten ein wichtiges Kriterium.

2.4 Datenkorrektheit

Die Untersuchung der Datenkorrektheit erfolgt oft auf Basis fachlicher Überlegungen und Validierungsregeln, häufig auch mit univariaten oberen und unteren Kontroll-Limits. In vielen Fällen ist die Überprüfung der Daten mit individuellen Limits hilfreich, um die Anzahl der Falsch-Positiven und Falsch-Negativen zu minimieren. Diese Limits werden üblicherweise abhängig von Eigenschaften (z. B. Geschlecht, Segment, Region) des individuellen Kunden oder Patienten berechnet.

2.5 Statistische Eigenschaften

Aus statistischer Sicht sind auch Faktoren wie die Korrelation der Variablen in der Datenmatrix, Variabilität und die Form der Verteilung wichtig, um die Datenqualität und die direkte Verwendbarkeit der Daten für die Analyse bewerten zu können.

3 Entdeckung und Behandlung von fehlenden Werten in Querschnittsdaten

3.1 Univariate Häufigkeiten

Fehlende Werte in Querschnittsdaten werden häufig in Form von univariaten Häufigkeiten dargestellt. Siehe Abbildung 2.

Obs	Variable	NumberMissing	Proportion_Missing	N
1	Alter	753	0.07	10303
2	EigenheimWert	945	0.09	10303
3	Einkommen	898	0.09	10303
4	EinkommensKlasse	359	0.03	10303
5	EurotaxKlasse	196	0.02	10303
6	EurotaxWert	244	0.02	10303
7	Fuehrerscheinentzug	151	0.01	10303
8	Geschlecht	792	0.08	10303
9	KundeSeit	924	0.09	10303
10	VermahnungsPunkte	124	0.01	10303

Abbildung 2: Univariate Darstellung der fehlenden Werte

Diese Darstellung beantwortet zwar Fragen wie: „Welche Variablen in meinen Daten leiden stark unter der „Fehlende-Werte Krankheit?“, betrachtet dies aber nur aus einer „Spalten-Perspektive“.

Es wird aber keine Antwort gegeben auf die Frage „Wie viele „Full-Records“ (=Records ohne Missing Values) im Datensatz enthalten sind, oder ob es ein Muster in der Struktur der fehlenden Daten gibt.

3.2 Struktur der fehlenden Daten

Eine bessere Aussage liefert hier das Profiling der Struktur der fehlenden Daten. Dazu wird für jede Beobachtung ein String mit 0 und 1 erzeugt, der anzeigt, ob die jeweilige Variable einen fehlenden Wert für diese Beobachtung enthält. Zum Beispiel bedeutet bei den drei Variablen Alter, Einkommen und Familienstand der String 010, dass Alter und Familienstand vorhanden sind, das Einkommen jedoch fehlt.

Diese Strings zeigen somit ein Muster der fehlenden Daten. Eine mögliche Darstellung ist ein sog. Tile-Chart (Kacheldiagramm) wie in Abbildung 3 dargestellt.

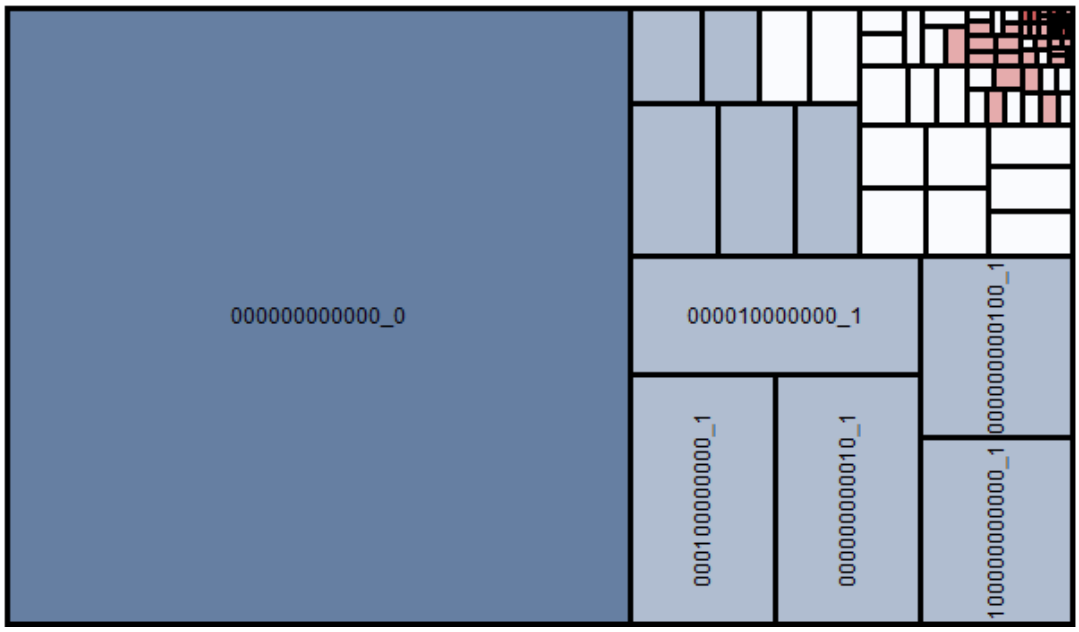

Abbildung 3: Profiling der Struktur der fehlenden Daten.

In dieser Graphik erkennen Sie den Anteil der „Full-Records“ von ca. 60 %, den Anteil der Beobachtungen mit jeweils einem fehlenden Wert und die kleine Gruppe von Beobachtungen mit einer höheren Anzahl von Missing Values in der rechten oberen Ecke.

3.3 Umsetzung in SAS und JMP

Die Möglichkeit, die fehlenden Werte so darzustellen, findet sich in JMP im Task „Struktur der fehlenden Daten“ , bzw. kann mit dem SAS Makro %MV_Profiling durchgeführt werden. Dieses Makro kann (mit anderen Data Quality Profiling Makros) heruntergeladen werden. Siehe dazu Kapitel 6.

```
%MV_Profiling(data=EM.KFZ_STORNO_DQ,
   vars= Alter AutoTyp AutoVerwendung  EigenheimWert Einkommen
   EinkommensKlasse EurotaxKlasse EurotaxWert Fuehrerscheinentzug
   Geschlecht KundeSeit VermahnungsPunkte );
```

3.4 Multivariate Analyse der Muster der fehlenden Werte

Das Makro %MV_Profiling erlaubt auch eine multivariate Analyse der Struktur der fehlenden Daten in Form einer Hauptkomponentenanalyse oder einer Variablen-Clusteranalyse. Die Ergebnisse können verwendet werden um zu beurteilen, welche Variablen sich „nahe“ sind aufgrund der Tatsache, dass Ihre Werte gemeinsam fehlen. Abbildung 4 zeigt ein graphisches Beispiel dazu.

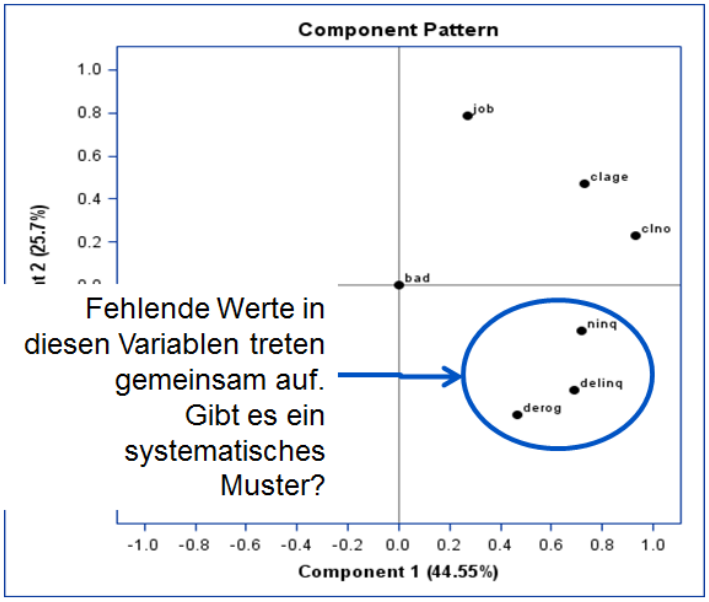

Abbildung 4: Hauptkomponentenanalyse zur Erkennung systematischer Muster in fehlenden Werten.

4 Entdeckung und Behandlung von fehlenden Werten in Längsschnittdaten

4.1 Profiling der Struktur von fehlenden Daten in Zeitreihen

Analog zum Beispiel der Querschnittsdaten, kann auch das Muster in Zeitreihendaten mit Hilfe von Strings dargestellt werden. Für jeden Zeitpunkt wird dabei der Wert 1, 0 oder X angehängt, wobei 1 für einen vorhandenen Wert, 0 für einen 0-Wert und X für einen fehlenden Wert steht. Auf diese Weise können die Häufigkeiten der Muster in den Zeitreihen dargestellt werden.

TS_Profile_Chain	Frequency	Percent
11111111111111111111111111111111111111111111111111111_54_0	18	39.13
111111111111111111111111111111111111111111111111111111111111_60_0	17	36.96
000000111111111111111111111111111111111111111111111111111111_60_0	5	10.87
111111001111111110000111111111111111111111111111111111000001_60_0	1	2.17
111111111111111111111110000000000001111111111111111111111111_60_0	1	2.17
111111111111111111111111001111111111111111111111111111111111_60_0	1	2.17
11111111111111111111111111111111111111111111111111111_53_0	1	2.17
111111111111111111111111111X1111111111111XX1X1XX11111XXXX_60_10	1	2.17
1111XX11111111111111111111111111111111X1X11XXXX11111XX111_60_10	1	2.17

Abbildung 5: Profiling des Fehlende Werte Musters von Zeitreihendaten

Am Beispiel von Abbildung 5 ist zu sehen, dass 18 Zeitreihen der Länge 54 durchgängig Werte besitzen ebenso 17 Zeitreihen der Länge 60. Es gibt einige Zeitreihen mit eingebetteten 0-Werten und Zeitreihen mit eingebetteten fehlenden Werte.

Abbildung 6 zeigt ein weiteres Beispiel, wie die Vollständigkeit einer Zeitreihe graphisch dargestellt werden kann. Es ist deutlich zu erkennen, dass es einen großen Block mit durchgängigen Zeitreihen gibt. Weiter gibt es mehr als die Hälfte der Zeitreihen, die erst zu einem späteren Zeitpunkt beginnen bzw. früher enden. Einige Zeitreihen haben eingebettete fehlende Werte. Der Vorteil dieser Art der Analyse und Darstellung ist, dass man auf einem Blick die Vollständigkeit der verfügbaren Analysendaten überblicken kann, was bei bloßer Betrachtung der Analysetabelle nicht möglich ist.

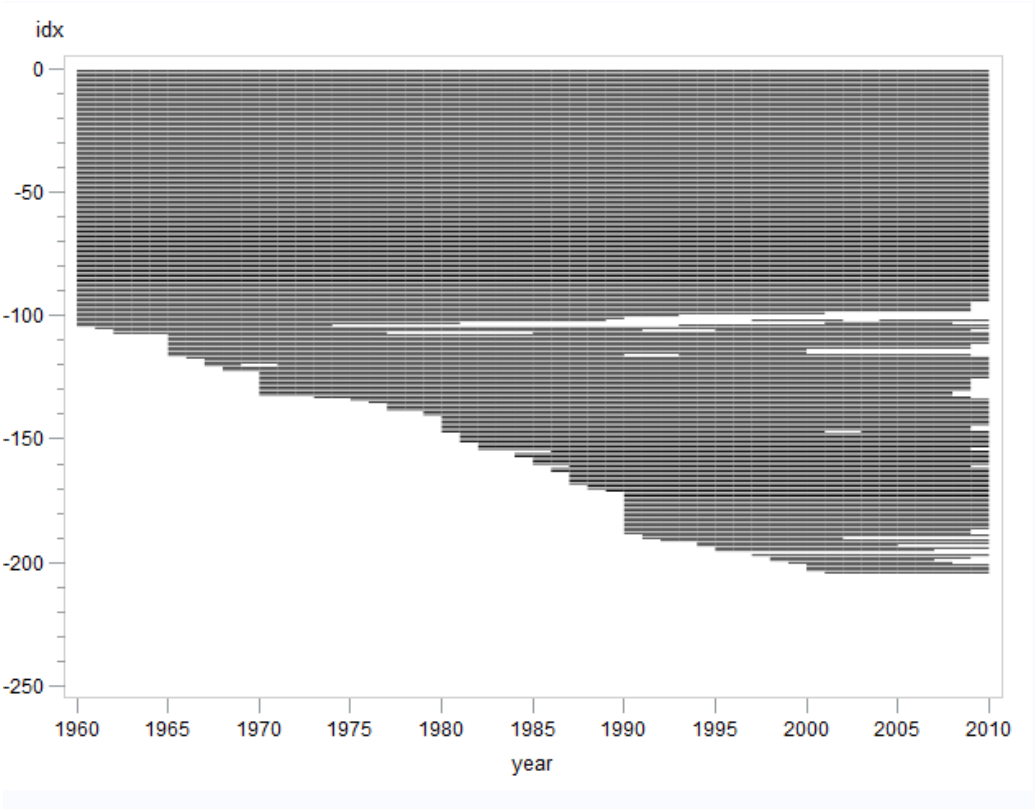

Abbildung 6: Plot der Vollständigkeit der Zeitreihen

Die Tabelle in Abbildung 5 und die Graphik in Abbildung 6 können mit dem Makro %PROFILE_TS_MV erstellt werden. (Details zur Verfügbarkeit siehe Kapitel 6.)

Codebeispiel für den Aufruf:

```
%Profile_TS_MV(data=tmp.gdp2, id=Country_name, date=Year, value=gdp,
               mv=(.,0), w=1, NMAX_TS=300);
```

4.2 Explizite und implizite fehlende Werte in Zeitreihendaten

Fehlende Werte in Zeitreihen können auf zwei Arten vorkommen. **Explizite fehlende Werte** sind durch einen fehlenden Wert (leeres Feld zu erkennen), wie in Zeile 6 in Abbildung 7 zu erkennen.

In dieser Tabelle befinden sich aber am Ende der Tabelle noch weitere fehlende Werte, die nicht auf den ersten Blick erkennbar sind, da es keinen fehlenden Wert im Feld AMOUNT gibt. Zwischen dem Monat September 2005 und Dezember 2005 fehlen beispielsweise 2 Beobachtungen, jene für den Oktober und für den November. In diesem Fall spricht man von **impliziten fehlenden Werten**.

PNR	date	amount
56	2004-02-01	48
56	2004-03-01	51
56	2004-04-01	42
56	2004-05-01	36
56	2004-06-01	6
56	2004-07-01	.
56	2004-08-01	48
56	2004-09-01	36
56	2004-10-01	66
56	2004-11-01	15
56	2004-12-01	33
58	2005-06-01	39
58	2005-07-01	63
58	2005-08-01	84
58	2005-09-01	18
58	2005-12-01	69
58	2006-03-01	0
58	2006-07-01	90
58	2006-10-01	57
58	2007-01-01	48

Abbildung 7: Zeitreihendaten mit explizt und implizt fehlenden Daten

Zur Erkennung und Ersetzung der fehlenden Werte kann die Prozedur TIMESERIES (als Bestandteil von SAS/ETS) verwendet werden. Das zugehörige Makro lautet %CHECK_TIMEID.

Ein Codebeispiel für die Prozedur TIMESERIES für die Aufdeckung von fehlenden Beobachtungen und die Ersetzung der Werte mit 0 zeigen die folgenden Programmzeilen:

```
PROC TIMESERIES DATA=air_missing OUT=timeid_inserted;
   ID date INTERVAL = month SETMISS=0;
   VAR air;
RUN;
```

4.3 Ersetzen von fehlenden Werte in Zeitreihen

Fehlende Werte in Zeitreihendaten können entweder wie im obigen Beispiel mit der Prozedur TIMESERIES durch Werte wie 0, Mittelwert, voriger/nächster/erster/letzter Wert ersetzt werden. Eine andere Möglichkeit bietet die Prozedur EXPAND, die ebenfalls Bestandteil von SAS/ETS ist, wo fehlende Werte durch SPLINE-Interpolationen ersetzt werden können.

Abbildung 8 zeigt ein Datenbeispiel mit fehlenden Werten in der AIR_MV Spalte.

	DATE	AIR	air_mv
1	JAN49	112	112
2	FEB49	118	118
3	MAR49	132	132
4	APR49	129	129
5	MAY49	121	.
6	JUN49	135	135
7	JUL49	148	.
8	AUG49	148	148
9	SEP49	136	136
10	OCT49	119	119
11	NOV49	104	.
12	DEC49	118	118
13	JAN50	115	115
14	FEB50	126	126
15	MAR50	141	141

Abbildung 8: Zeitreihendaten mit fehlenden Werten

Die ersetzten Werte sind in Abbildung 9 in der Spalte AIR_EXPAND zu sehen.

	date	air	air_mv	air_expand
1	JAN49	112	112	112
2	FEB49	118	118	118
3	MAR49	132	132	132
4	APR49	129	129	129
5	MAY49	121	.	128.29783049
6	JUN49	135	135	135
7	JUL49	148	.	144.73734152
8	AUG49	148	148	148
9	SEP49	136	136	136
10	OCT49	119	119	119
11	NOV49	104	.	116.19900978
12	DEC49	118	118	118
13	JAN50	115	115	115
14	FEB50	126	126	126
15	MAR50	141	141	141
16	APR50	135	135	135
17	MAY50	125	125	125

Abbildung 9: Zeitreihendaten mit imputierten Werten

PROC EXPAND kann nun verwendet werden, um diese fehlenden Werte zu ersetzen, wie folgenden Code Beispiel zeigt:

```
PROC EXPAND DATA = AIR_MISSING
              OUT = AIR_EXPAND;
   CONVERT AIR_MV = AIR_EXPAND;
   ID DATE;
RUN;
```

5 Zusammenfassung

Datenaufbereitung für Analytik und die Betrachtung der Datenqualität unter analytischen Gesichtspunkten sind eine wichtige Voraussetzung für statistische Analysen. Diese beiden Elemente bilden die Brücke zwischen einer fachlichen Fragestellung und der statistischen Auswertung.

In dieser Arbeit wurden ausgewählte Methoden des statistischen Profilings der Analysedaten gezeigt. Weitere Methoden finden sich im Buch „Data Quality for Analytics“ [1] bzw. in Präsentationen, die auf der Download-Seite zu diesem Buch, vgl. Kapitel 6, aufgeführt werden.

6 Downloads

Daten und Programme zu „Data Quality for Analytics Using SAS“ finden Sie unter:
http://www.sascommunity.org/wiki/Data_Quality_for_Analytics_--_Download_Page
und zu
„Data Preparation for Analytics Using SAS“ unter:
http://www.sascommunity.org/wiki/Data_Preparation_for_Analytics

Bei Fragen und Kommentaren können Sie sich gerne an den Autor wenden bzw. die Seite http://www.sascommunity.org/wiki/Gerhard_Svolba besuchen.

Literatur

[1] Svolba, Gerhard: Data Quality for Analytics Using SAS, SAS Press 2012

[2] Svolba, Gerhard: Data Preparation for Analytics Using SAS; SAS Press 2006

[3] Svolba, Gerhard: Data Preparation for Data Mining, KSFE 2008