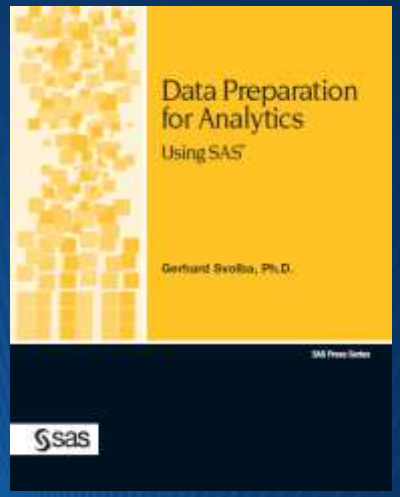
WBS Herbstseminar:
„Buchpräsentationen aus dem Bereich der Statistik“



DATA QUALITY FOR ANALYTICS USING SAS

sas
THE POWER TO KNOW.

DR. GERHARD SVOLBA
WIEN, 21. NOVEMBER 2013

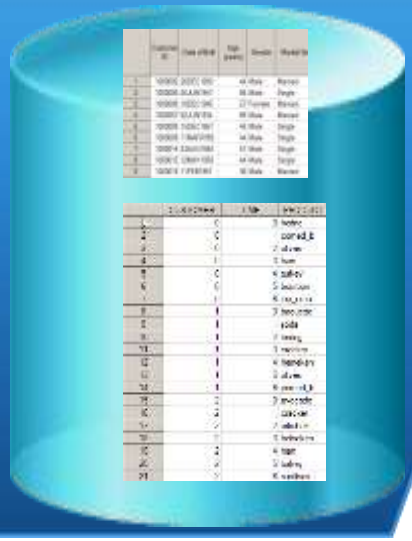
Scope von Datenaufbereitung und Datenqualität für Analytik

Fachliche Fragestellung

Datenaufbereitung Datenqualität

1. Review der fachlichen Fragestellung
2. Definition der Datenqualitätskriterien
3. Erzeugen eines Data Marts aus den Inputdaten
4. Profiling des Datenqualitäts-Status
5. Entscheidung über den Datenqualitätsstatus
6. Aufbereitung der Daten für die analytische Verwendung

Finaler Analysemart



Das gesamte Ökosystem (Entscheidungen, Kriterien, Datenaufbereitungsschritte, fachliche Überlegungen) das zwischen der FACHLICHEN FRAGESTELLUNG und dem FINALEN ANALYTISCHEN MART liegt.

„ÜBER DEN VORTRAGENDEN“

- Produktverantwortlicher für die SAS Analytik Produkte
- Analytik Solution Architekt bei SAS Austria
- Buchautor bei SAS Press
- Begeisterter Segler

KLEINE SEGEL- UND REGATTAKUNDE (1)

Gemeinsamer Start gegen den Wind entlang einer Linie zwischen Startschiff und Boje

KLEINE SEGEL- UND REGATTAKUNDE (2)

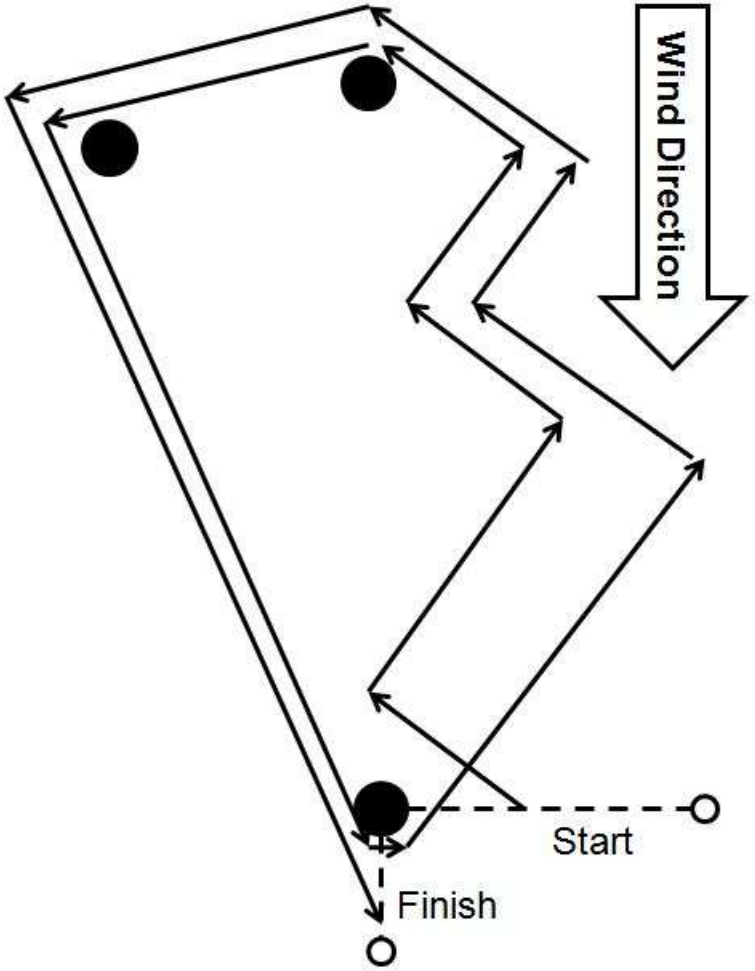
Aufkreuzen gegen den Wind.
Mehrmaliges Wenden, da die nächste Boje nicht am direkten Weg erreicht werden kann.

KLEINE SEGEL- UND REGATTAKUNDE (3)

Nach dem Runden der Boje: Segeln mit achterlichem Wind und Spinnaker zur nächsten Boje

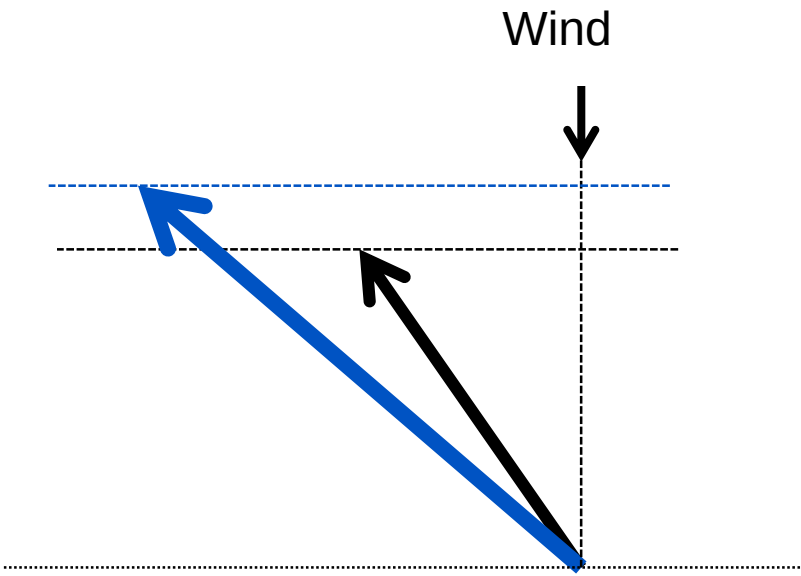
KLEINE SEGEL- UND REGATTAKUNDE (4)

Skizze einer Regattabahn mit 3 Bojen.



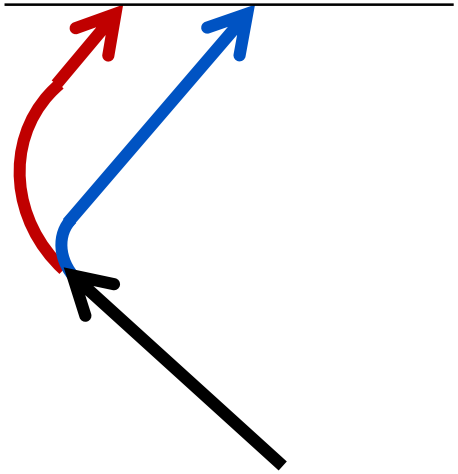
HIER KOMMT DER STATISTIKER INS SPIEL!
ANALYSEFRAGEN ZUR OPTIMIERUNG DER SEGELTAKTIK 1

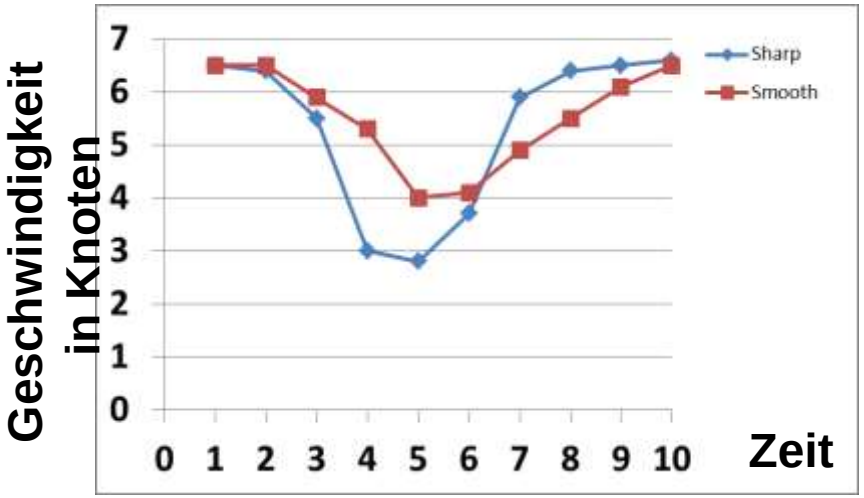
- Welcher Kurswinkel zum „wahren“ Wind soll gesegelt werden?
  - Je spitzer der Winkel, umso direkter der Kurs, umso kürzer die Strecke die gesegelt werden muss.
  - Abhängigkeit von der Windstärke und von der Besegelung



HIER KOMMT DER STATISTIKER INS SPIEL!
ANALYSEFRAGEN ZUR OPTIMIERUNG DER SEGELTAKTIK 2

- Wie soll gewendet werden?
  - **Rasch, zackig**: damit das Boot gleich wieder auf neuen Kurs kommt und Fahrt aufnimmt?
  - **Rund, fließend**: damit das Boot in der Wende möglichst wenig Fahrt verliert?





VERFÜGBARE DATEN FÜR DIE SEGEL-ANALYSEN: „GPS-TRACKPOINT DATEN“

- Long/Lat Position
- Kurs (Kompass)
- Geschwindigkeit


```
 <MetadataTag name="SailorName" value="xxxx" />
  </MetadataTags>
  <CapturedTrack name="090521_131637" downloadedOn="2009-05-25T18:23:46.25+02:00"
numberTrkpts="8680">
    <MinLatitude>47.773464202880859</MinLatitude>
    <MaxLatitude>47.804649353027344</MaxLatitude>
    <MinLongitude>16.698064804077148</MinLongitude>
    <MaxLongitude>16.74091911315918</MaxLongitude>
    <DeviceInfo ftdiSerialNumber="VTQURQX9" />
    <SailorInfo firstName="xxxx" lastName="yyyy" yachtClub="zzzz" />
    <BoatInfo boatName="wwww" sailNumber="0000" boatClass="Unknown" hullNumber="0" />
    <Trackpoints>
      <Trackpoint dateTime="2009-05-21T13:49:24+02:00" heading="68.43" speed="5.906" latitude="47.792442321777344"
longitude="16.727603912353516" />
      <Trackpoint dateTime="2009-05-21T13:49:26+02:00" heading="59.38" speed="5.795" latitude="47.7924690246582"
longitude="16.727682113647461" />
      <Trackpoint dateTime="2009-05-21T13:49:28+02:00" heading="65.41" speed="6.524" latitude="47.792495727539062"
longitude="16.727762222290039" />
      <Trackpoint dateTime="2009-05-21T13:49:30+02:00" heading="62.2" speed="6.631" latitude="47.792518615722656"
longitude="16.727849960327148" />
      <Trackpoint dateTime="2009-05-21T13:49:32+02:00" heading="56.24" speed="6.551" latitude="47.792549133300781"
longitude="16.727928161621094" />
      <Trackpoint dateTime="2009-05-21T13:49:34+02:00" heading="60.56" speed="5.978" latitude="47.792579650878906"
longitude="16.728004455566406" />
      <Trackpoint dateTime="2009-05-21T13:49:36+02:00" heading="61.57" speed="7.003" latitude="47.792606353759766"
longitude="16.728090286254883" />
      <Trackpoint dateTime="2009-05-21T13:49:38+02:00" heading="52.03" speed="7.126" latitude="47.792636871337891"
longitude="16.728176116943359" />
```

VERFÜGBARE DATEN FÜR DIE SEGEL-ANALYSEN: „HÄNDISCHE AUFZEICHNUNGEN“

- Zusammensetzung der Crew
- Segelgröße und Segeltyp
- Windstärke und Windrichtung
- Platzierung in der Wettfahrt
- Sonstige Kommentare

Datum	Regatta	Wettfahrt	Steuermann	Mittelmann	Spimann	Grossegel	Vorsegel	Spi	Spi gesetzt	Windstärke	Windrichtung
21.05.2009	Ruster Segeltage	1	Günter	Christian	Gerhard	Binder	2	Rot	1	3-4	S
21.05.2009	Ruster Segeltage	2	Günter	Christian	Gerhard	Binder	2	Rot	1	3-4	S
21.05.2009	Ruster Segeltage	3	Günter	Christian	Gerhard	Binder	2	Rot	1	3-4	S
22.05.2009	Ruster Segeltage	4	Günter	Christian	Gerhard	Binder	1	Rot	1	1-2	NW
23.05.2009	Ruster Segeltage	5	Günter	Christian	Gerhard	Binder	3	Rot	1	2-3	NW
23.05.2009	Ruster Segeltage	6	Günter	Christian	Gerhard	Binder	2	Rot	1	2-3	NW
23.05.2009	Ruster Segeltage	7	Günter	Christian	Gerhard	Binder	2	Rot	1	2-3	NW
20.06.2009	Blaues Band	1	Günter	Karl	Gerhard	Binder?	2	Rot	1	4-5	NW
27.06.2009	3 Insel	1	Günter	Karl	Gerhard	Binder	1	Rot	1	2-3	NW
27.06.2009	3 Insel	2	Günter	Karl	Gerhard	Binder	1	Rot	1	2	NW
27.06.2009	3 Insel	3	Günter	Karl	Gerhard	Binder	1	Rot	1	2	NW
28.06.2009	3 Insel	4	Günter	Karl	Gerhard	Binder	1	Rot	1	1	NW
28.06.2009	3 Insel	5	Günter	Karl	Gerhard	Binder	1	Rot	1	1	NW
25.07.2009	CBS-Cup	1	Günter	Karl	Gerhard	Binder	3	Rot	0	6	NW
26.07.2009	CBS-Cup	2	Günter	Karl	Gerhard	Binder	2	Rot	1	2-3	NW
26.07.2009	CBS-Cup	3	Günter	Karl	Gerhard	Binder	2	Rot	1	2-3	NW
26.07.2009	CBS-Cup	4	Günter	Karl	Gerhard	Binder	1	Rot	1	2	NW
19.09.2009	Absegeln	1	Günter	Michael Reite	Marlene + M	Binder	1	Rot	1	1	NW

SEGELWELT
≈
GESCHÄFTSWELT

KATEGORISIERUNG DER DATENQUALITÄTSPROBLEME

- Ausfall des GPS-Device wegen niedriger Temperaturen und schlechter Batterien
- Trimm-Einstellungen des Bootes wurden nicht dokumentiert
- Händische Aufzeichnungen: teilweise lückenhaft, oft erst im nachhinein erstellt
- In seltenen Fällen: Long/Lat Positionierung zeitlich verzögert → falsche Berechnung der Geschwindigkeit
- Datentransfer:
  - GPS-Device → (XML) → PC
  - XML/Text → SAS
- Nur 97 „Wenden“ im ersten Jahr in den Daten dokumentiert

Data Completeness

Data Correctness

Data Quantity

SEGELWELT
≈
GESCHÄFTSWELT

KATEGORISIERUNG DER DATENQUALITÄTSPROBLEME

- Data Cleaning: Daten-Sammlung vom Einschalten bis zum Ausschalten des Geräts

Data Usability

- Keine GPS Trackpoint-Daten von anderen Segelbooten verfügbar
- Windstärke und Windrichtungsdaten werden am Boot nicht erfasst
- Externe Daten: Mess-Station am Land, andere Zeitintervalle, historische Verfügbarkeit

Data Availability

SEGELWELT

≈

GESCHÄFTSWELT

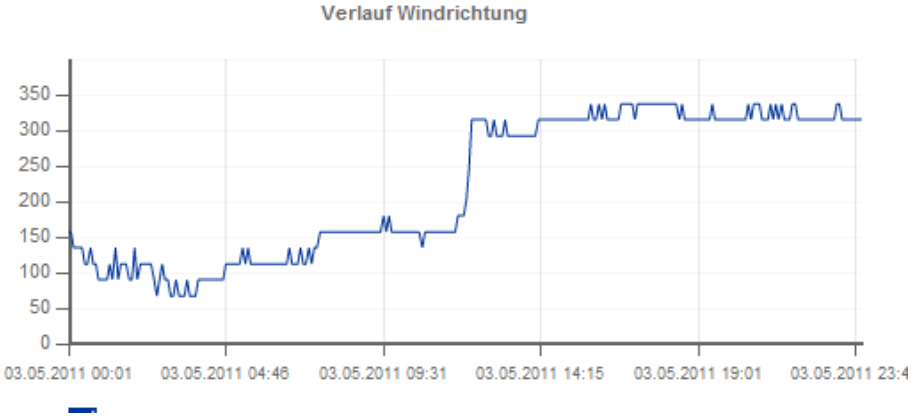
VERFÜGBARKEIT VS. VERWENDBARKEIT VON DATEN AM BEISPIEL VON WETTER- UND WINDDATEN

▶ Ruster Bucht / Neusiedler See

▶ Wetterstation Rust 10.05.2011, 12:10 Uhr

Windstärke	1 Bft / 2,7 kts
Windrichtung aktuell	O-NO / 67 °
Hauptwindrichtung	W
Lufttemperatur	19,7 °C
Wassertemperatur	16,1 °C
Wind-Chill	19,7 °C
Taupunkt	8,8 °C
Wärmebelastung	leichte Wärmebelastung (15)
Luftdruck	1026,5 hPa
Relative Luftfeuchte	49 %
Niederschlag 1h/24h	0,0 l/m² / 0,2 l/m²

Quelle: www.byc.at

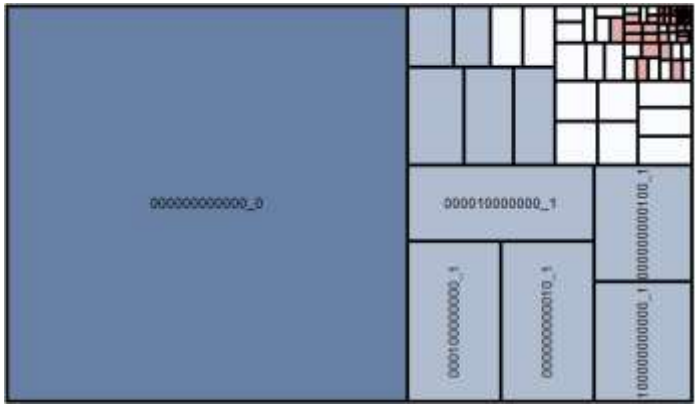


DQ-KRITERIEN FÜR ANALYTIK

DIE SICHTWEISE VON ANALYTIK AUF DATENQUALITÄT

- **Datenverfügbarkeit**
  - Aktuelle Daten, historische Daten, „Snapshots“ zu Zeitpunkten in der Vergangenheit
  - Gewährleistung der fortlaufenden Verfügbarkeit
  - Granularitätslevel: Aggregate oder Einzeldaten
- **Datenmenge**
  - Anzahl der Analyse-Subjekte, Beobachtungszeitraum, Anzahl der Ereignisse
- **Datenvollständigkeit**
  - Zufällige oder systematische fehlende Werte, Muster
  - Aufwand der Vervollständigung der Daten
- **Datenkorrektheit**
  - Univariate und multivariate Plausibilitätskontrollen
- **Statistische Eigenschaften**
  - Korrelation, Variabilität und Verteilung

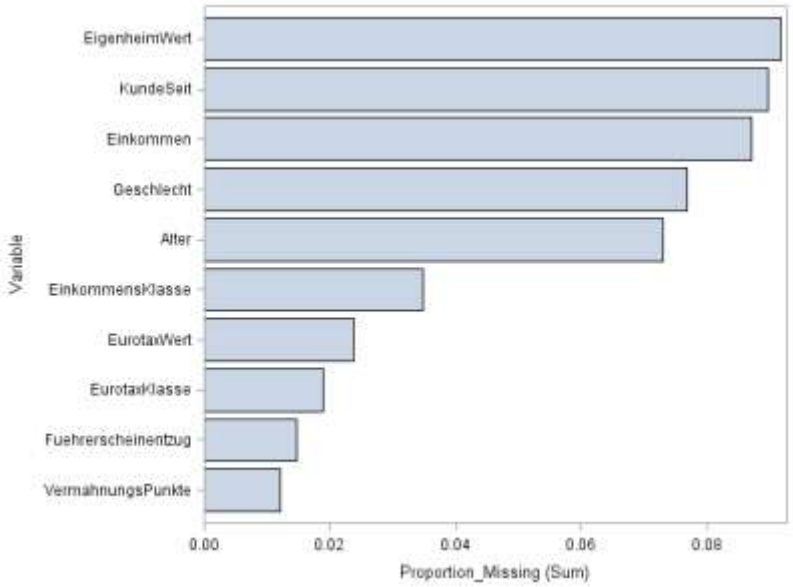
PROFILING VON ANALYTISCHER DATENQUALITÄT



FEHLENDE WERTE IN QUERSCHNITTS-DATEN

UNIVARIATE ANALYSE DER HÄUFIGKEITEN

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

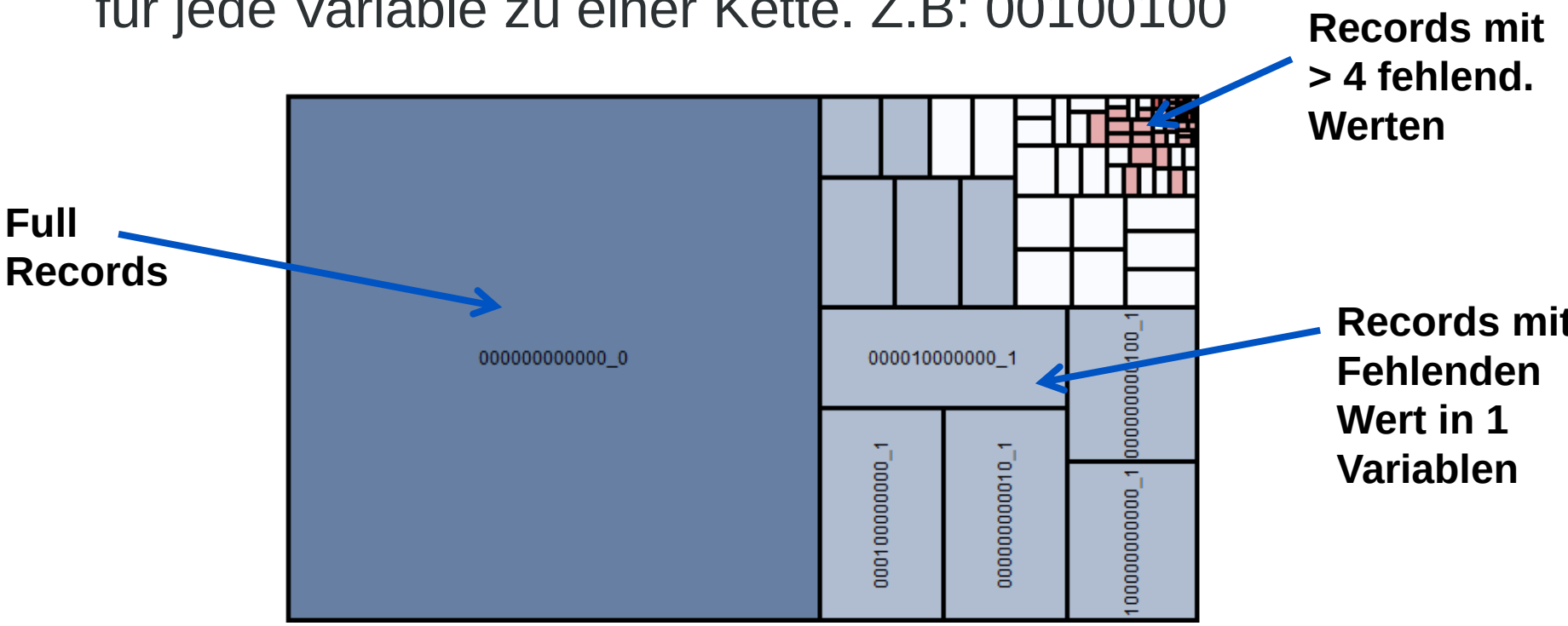


- Welche Variablen in meinen Daten leiden unter der „Fehlende-Werte Krankheit“?
- Betrachtung aber nur aus einer „Spalten-Perspektive“
- Nicht: Wie viele „Full-Records“ habe ich?
- Nicht: Gibt es ein Muster in der Struktur der fehlenden Daten?

FEHLENDE WERTE IN QUERSCHNITTS-DATEN

MULTIVARIATE ANALYSE BRINGT MEHR EINBLICK

- Zusammenhängen der „Fehlender Wert“-Indikatoren für jede Variable zu einer Kette. Z.B: 00100100




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

FEHLENDE WERTE IN LÄNGS-SCHNITTDATEN

PROFILING DER STRUKTUR VON FEHLENDEN WERTEN UND NULL-WERTEN IN LÄNGSSCHNITTDATEN

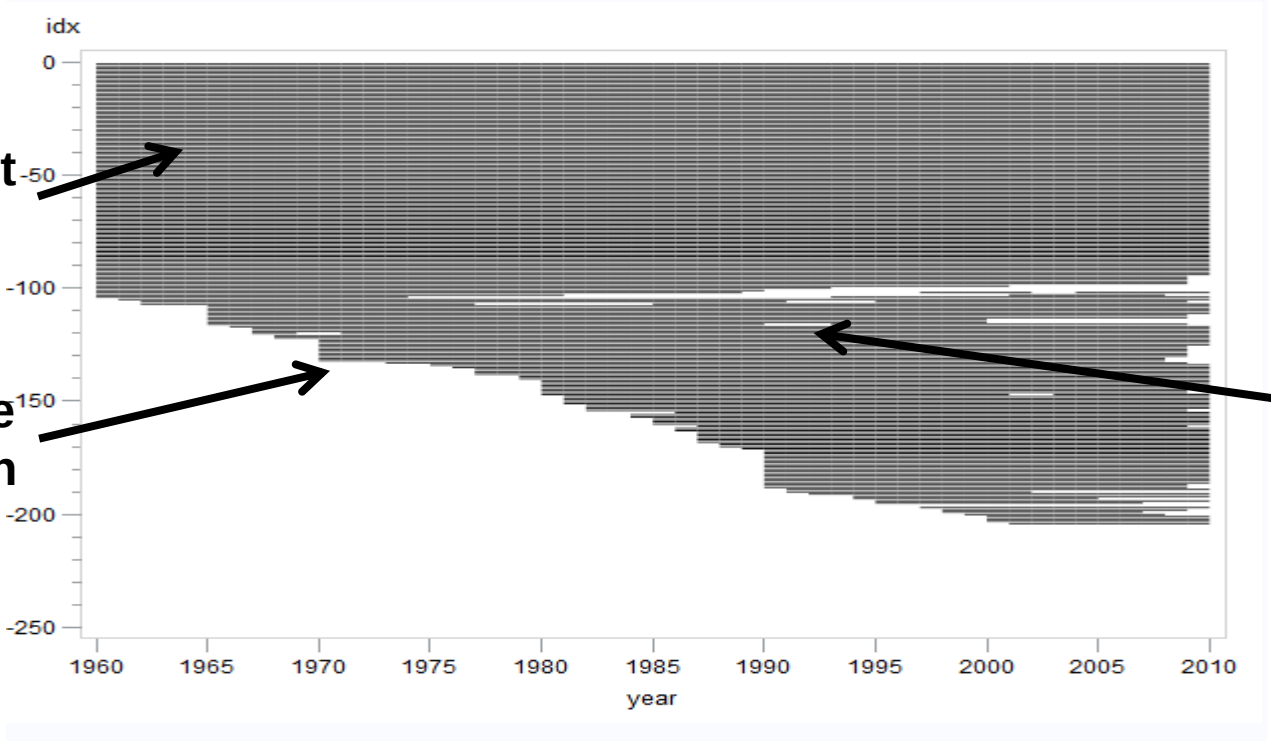
TS_Profile_Chain	Frequency	Percent
11111111111111111111111111111111111111111111111111111111_54_0	18	39.13
111111111111111111111111111111111111111111111111111111111111_60_0	17	36.96
000000111111111111111111111111111111111111111111111111111111_60_0	5	10.87
111111001111111111000011111111111111111111111111111111111000001_60_0	1	2.17
111111111111111111111111000000000000011111111111111111111111_60_0	1	2.17
111111111111111111111111001111111111111111111111111111111111_60_0	1	2.17
111111111111111111111111111111111111111111111111111111111111_53_0	1	2.17
111111111111111111111111111111X1111111111111XX1X1XX11111XXXX_60_10	1	2.17
1111XX11111111111111111111111111111111111X1X1XXXX11111XX111_60_10	1	2.17

0 Value

Fehlender Wert

FEHLENDE WERTE IN LÄNGS-SCHNITTDATEN

PROFILING DER STRUKTUR VON FEHLENDEN WERTEN UND NULL-WERTEN IN LÄNGSSCHNITTDATEN



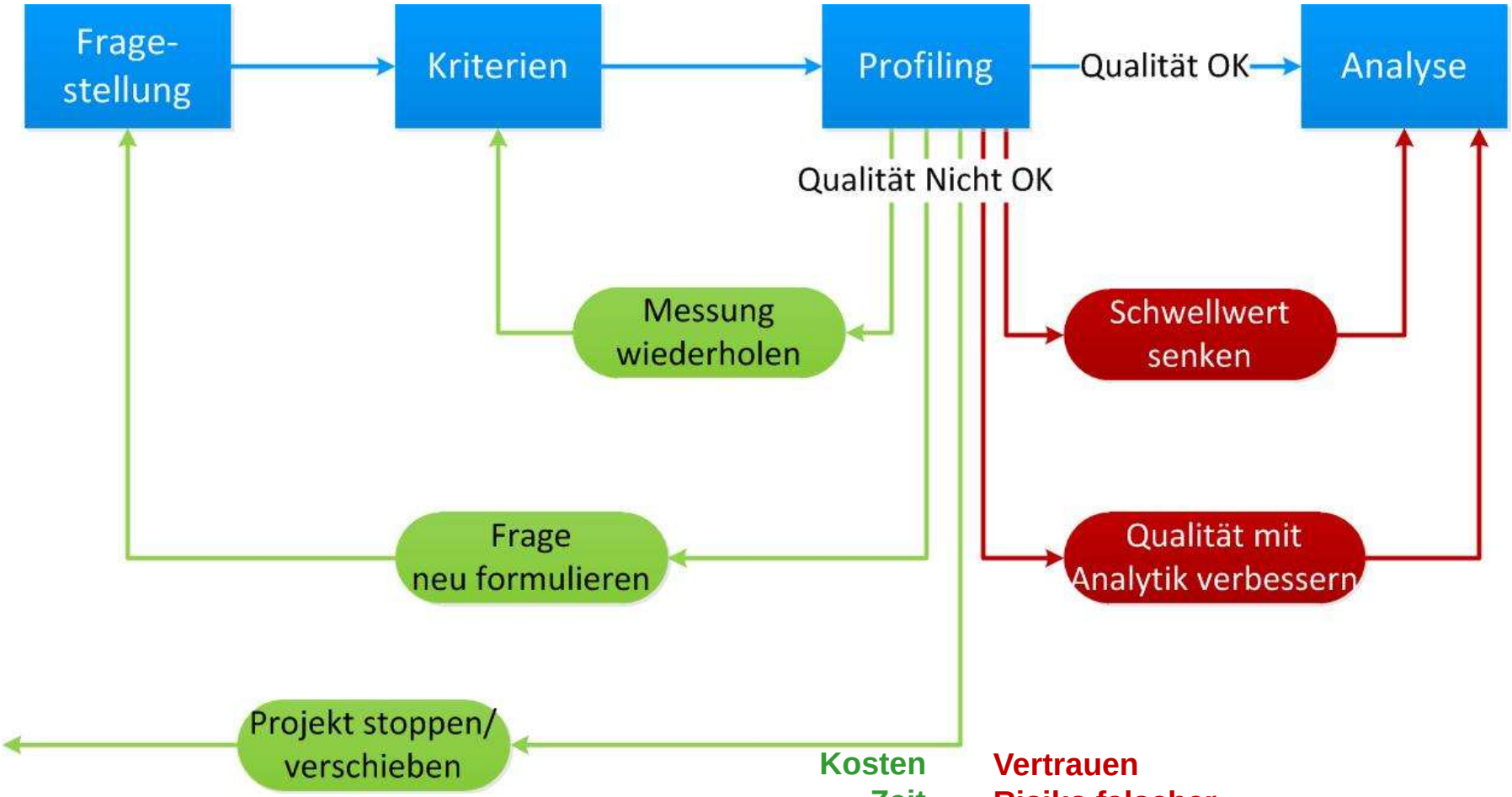
Zeitreihen mit voller Länge

Zeitreihen die später starten

Zeitreihen mit eingebetteten fehlenden Werten

AUSSCHNITT AUS DEM ANALYSEPROZESS

DAS SIND IHRE OPTIONEN, WENN SIE ERKENNEN, DASS DIE DATENQUALITÄT SCHLECHT IST



Kosten
Zeit
Verzögerungen
Keine Resultate

Vertrauen
Risiko falscher
Entscheidungen
Insignifikanz

SIMULATION DER KONSEQUENZEN SCHLECHTER DATENQUALITÄT IM DATA MINING

SIMULATIONS-IDEE

DIE KONSEQUENZEN DER FOLGENDEN EFFEKTE WURDEN UNTERSUCHT:

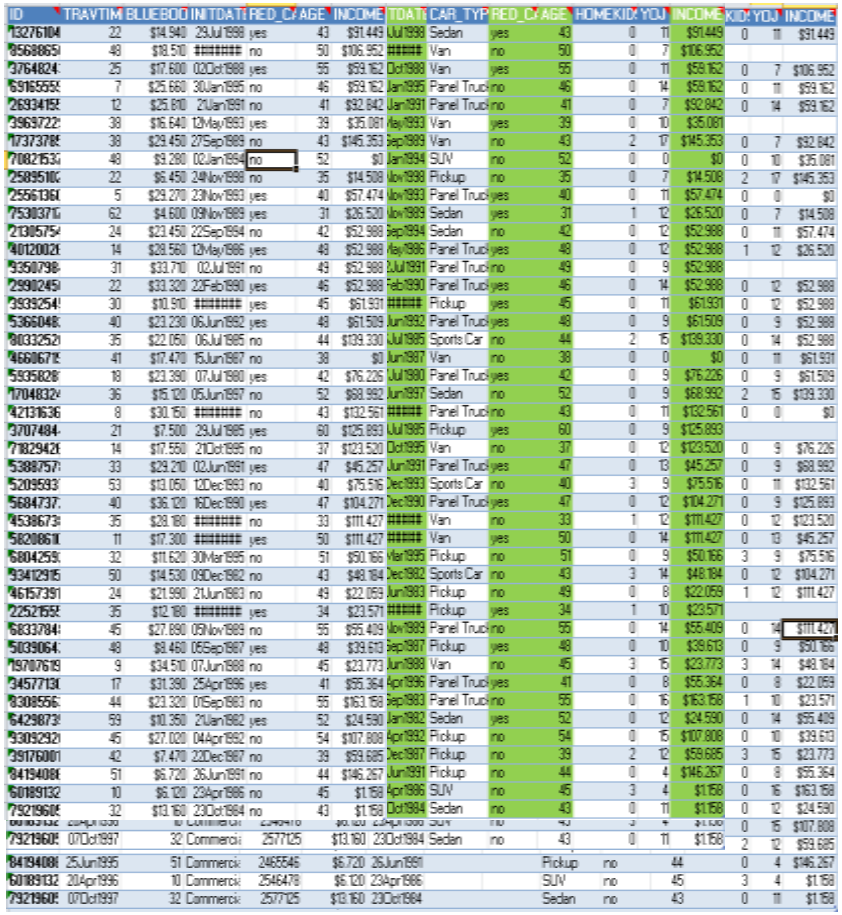
- Wie beeinflussen fehlende Werte die Vorhersagekraft? (SIM_1)
- Variation der Anzahl der verfügbaren Beobachtungen und Ereignisse (SIM_2)

- Andere Fragen / Simulationen
  - Entfernen der wichtigsten Variablen aus dem Input-Set
  - Einführung zufälliger und systematischer Verzerrungen in der Input- und Zielgröße im Predictive Modeling
  - Auswirkung von zufälligen und systematischen fehlenden Werte und Fehlern in der Zeitreihen Prognose
  - Schrittweise Erhöhung der verfügbaren Länge der Datenhistorie

SIMULATIONS-IDEE

ECHTDATEN WURDEN FÜR DIE SIMULATIONSSTUDIEN VERWENDET

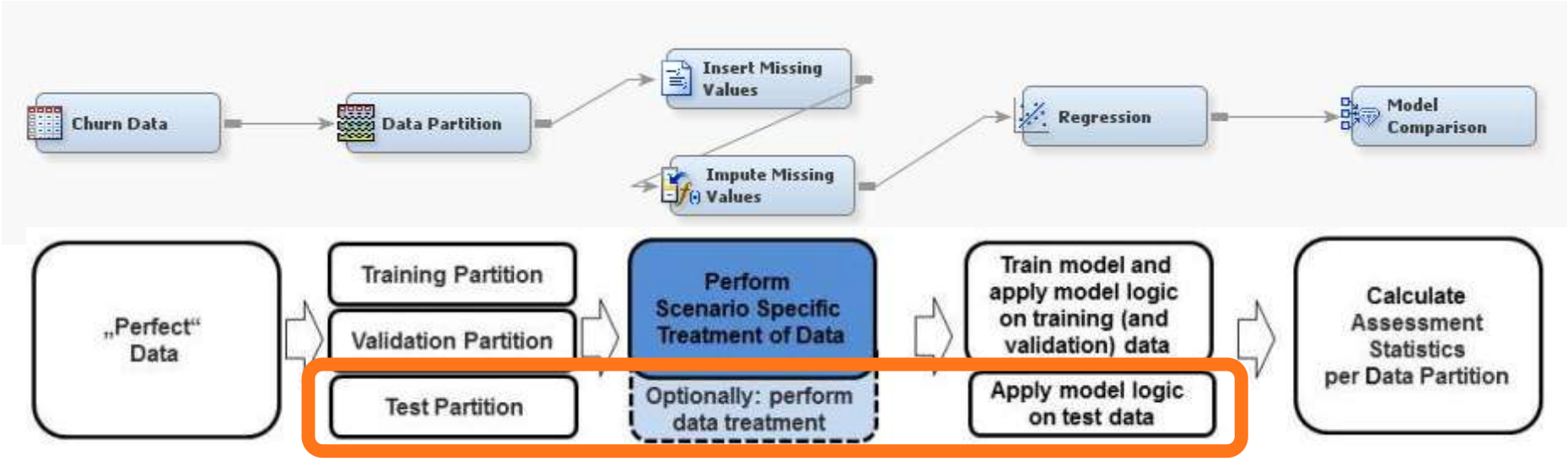
- 4 Echt-Datensätze aus unterschiedlichen Branchen mit einem binären Target wurden verwendet
- Variablen mit > 5 % fehlenden Werten wurden entfernt
- Bei Variablen mit ≤ 5 % fehlenden Werten wurden Beobachtungen mit Missings gefilert
- Durchlauf multipler Modell-Zyklen um ein stabiles Modell mit guter Vorhersagekraft zu bekommen



ID	TRAVTIM	BLUEBOO	INTDATI	RED_CA	AGE	INCOME	JOBTI	CAR_TYP	RED_CA	AGE	HOMEKI	KIDS	YOJ	INCOME	KIDS	YOJ	INCOME
13276104	22	$14,940	29Jul1998	yes	43	$31,449	Jul1998	Sedan	yes	43	0	11	$31,449	0	11	$31,449	
95688651	48	$18,510	#######	no	50	$106,952	#####	Van	no	50	0	7	$106,952				
3764824	25	$17,600	02Oct1988	yes	55	$59,162	Oct1988	Van	yes	55	0	11	$59,162	0	7	$106,952	
63165551	7	$25,660	30Jan1995	no	46	$59,162	Jan1995	Panel Truck	no	46	0	14	$59,162	0	11	$59,162	
26934151	12	$25,810	21Jan1991	no	41	$32,842	Jan1991	Panel Truck	no	41	0	7	$32,842	0	14	$59,162	
39637221	38	$16,640	12May1993	yes	39	$35,081	May1993	Van	yes	39	0	10	$35,081				
17373781	38	$28,450	27Sep1989	no	43	$145,353	Sep1989	Van	no	43	2	17	$145,353	0	7	$32,842	
70821531	48	$3,280	02Jan1994	no	52	$0	Jan1994	SUV	no	52	0	0	$0	0	10	$35,081	
25895101	22	$6,450	24Nov1998	no	35	$14,508	Nov1998	Pickup	no	35	0	7	$14,508	2	17	$145,353	
25561361	5	$29,270	23Nov1993	yes	40	$57,474	Nov1993	Panel Truck	yes	40	0	11	$57,474	0	0	$0	
75303711	62	$4,600	09Nov1989	yes	31	$26,520	Nov1989	Sedan	yes	31	1	12	$26,520	0	7	$14,508	
21305751	24	$23,450	22Sep1994	no	42	$52,988	Sep1994	Sedan	no	42	0	12	$52,988	0	11	$57,474	
40120021	14	$28,560	12May1986	yes	48	$52,988	May1986	Panel Truck	yes	48	0	12	$52,988	1	12	$26,520	
93507981	31	$33,710	02Jul1991	no	49	$52,988	Jul1991	Panel Truck	no	49	0	9	$52,988				
29902451	22	$33,320	23Feb1990	yes	46	$52,988	Feb1990	Panel Truck	yes	46	0	14	$52,988	0	12	$52,988	
39332541	30	$10,910	#######	yes	45	$61,931	#####	Pickup	yes	45	0	11	$61,931	0	12	$52,988	
53660481	40	$23,230	06Jun1992	yes	48	$61,509	Jun1992	Panel Truck	yes	48	0	9	$61,509	0	9	$52,988	
80332521	35	$22,050	06Jul1985	no	44	$139,330	Jul1985	Sports Car	no	44	2	15	$139,330	0	14	$52,988	
46606711	41	$17,470	15Jun1987	no	38	$0	Jun1987	Van	no	38	0	0	$0	0	11	$61,931	
59358281	18	$23,390	07Jul1990	yes	42	$76,226	Jul1990	Panel Truck	yes	42	0	9	$76,226	0	9	$61,509	
17048321	36	$15,120	05Jun1997	no	52	$68,992	Jun1997	Sedan	no	52	0	9	$68,992	2	15	$139,330	
42131631	8	$30,150	#######	no	43	$132,561	#####	Panel Truck	no	43	0	11	$132,561	0	0	$0	
37074841	21	$7,500	29Jul1985	yes	60	$125,893	Jul1985	Pickup	yes	60	0	9	$125,893				
71829421	14	$17,550	21Oct1995	no	37	$123,520	Oct1995	Van	no	37	0	12	$123,520	0	9	$76,226	
53887571	33	$29,270	02Jun1991	yes	47	$45,257	Jun1991	Panel Truck	yes	47	0	13	$45,257	0	9	$68,992	
52095931	53	$13,050	12Dec1993	no	40	$75,516	Dec1993	Sports Car	no	40	3	9	$75,516	0	11	$132,561	
56847371	40	$36,120	16Dec1990	yes	47	$104,271	Dec1990	Panel Truck	yes	47	0	12	$104,271	0	9	$125,893	
45386731	35	$28,180	#######	no	33	$111,427	#####	Van	no	33	1	12	$111,427	0	12	$123,520	
58208611	11	$17,300	#######	yes	50	$111,427	#####	Van	yes	50	0	14	$111,427	0	13	$45,257	
58042591	32	$11,620	30Mar1995	no	51	$50,166	Mar1995	Pickup	no	51	0	9	$50,166	3	9	$75,516	
93412911	50	$14,530	09Dec1982	no	43	$48,184	Dec1982	Sports Car	no	43	3	14	$48,184	0	12	$104,271	
46157391	24	$21,990	21Jun1983	no	49	$22,059	Jun1983	Pickup	no	49	0	8	$22,059	1	12	$111,427	
22521551	35	$12,180	#######	yes	34	$23,571	#####	Pickup	yes	34	1	10	$23,571				
58337841	45	$27,890	09Nov1989	no	55	$55,409	Nov1989	Panel Truck	no	55	0	14	$55,409	0	14	$111,427	
50390641	48	$8,460	05Sep1987	yes	48	$39,613	Sep1987	Pickup	yes	48	0	10	$39,613	0	9	$50,166	
19707619	9	$34,510	07Jun1988	no	45	$23,773	Jun1988	Van	no	45	3	15	$23,773	3	14	$48,184	
34577131	17	$31,390	25Apr1996	yes	41	$55,364	Apr1996	Panel Truck	yes	41	0	8	$55,364	0	8	$22,059	
83085561	44	$23,320	01Sep1983	no	55	$163,158	Sep1983	Panel Truck	no	55	0	16	$163,158	1	10	$23,571	
54298731	59	$10,350	21Jan1982	yes	52	$24,590	Jan1982	Sedan	yes	52	0	12	$24,590	0	14	$55,409	
93052921	45	$27,020	04Apr1992	no	54	$107,808	Apr1992	Pickup	no	54	0	15	$107,808	0	10	$39,613	
29176001	42	$7,470	22Dec1987	no	39	$59,685	Dec1987	Pickup	no	39	2	12	$59,685	3	15	$23,773	
84194081	51	$6,720	26Jun1991	no	44	$146,267	Jun1991	Pickup	no	44	0	4	$146,267	0	8	$55,364	
60189132	10	$6,120	23Apr1986	no	45	$1,158	Apr1986	SUV	no	45	3	4	$1,158	0	16	$163,158	
79219601	32	$13,160	23Oct1984	no	43	$1,158	Oct1984	Sedan	no	43	0	11	$1,158	0	12	$24,590	
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	15	$107,808	
79219601	07Oct1997	32	Commerci	2577125		$13,160	23Oct1984	Sedan	no	43	0	11	$1,158	2	12	$59,685	
84194081	26Jun1991	51	Commerci	2465546		$6,720	26Jun1991			Pickup	no	44		0	4	$146,267	
60189132	23Apr1986	10	Commerci	2546478		$6,120	23Apr1986			SUV	no	45		3	4	$1,158	
79219601	07Oct1997	32	Commerci	2577125		$13,160	23Oct1984			Sedan	no	43		0	11	$1,158	

SIMULATIONS-IDEE

SIMULATIONSSTUDIEN HELFEN DIE KONSEQUENZEN SCHLECHTER DATENQUALITÄT ZU QUANTIFIZIEREN



„Unberührte“ Testdata werden als „Scoring Daten“ verwendet

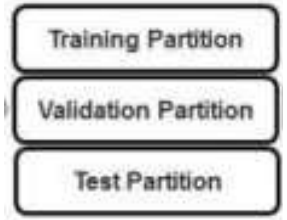
- Beobachtungen löschen
- Fehlende Werte einfügen und ersetzen

Verwenden der eingefrorenen Variablen-Liste auf die Scenario-Daten

SIMULATIONS ERGEBNISSE FEHLENDE WERTE

PROZESS FÜR DIE SZENARIOS MIT FEHLENDEN WERTEN

- Für einen bestimmten Anteil von Beobachtungen (10%, ...)
  - Setzen von Intervall-skalierten und nominalen Input-Variablen auf fehlend
    - » Zufällige Auswahl
    - » Systematische Auswahl auf Basis von Segmenten
  - Ersetzen der fehlende Werte durch den Impute Node im SAS Enterprise Miner
- Trainieren des Modells mit dem eingefrorenen Set an Variablen
- Optional: Anwenden der "Behandlung" auch für Scoring-Daten
- Bewerten der Modellqualität

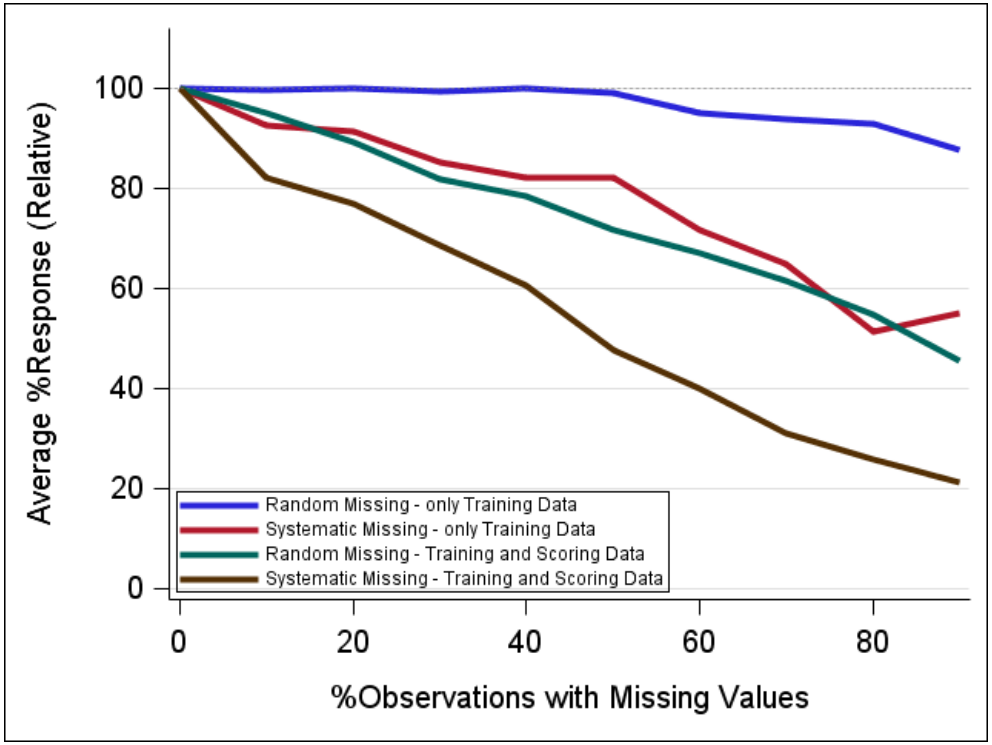


ID	TRAVTIM	BLUEBOO	INITDATE	RED_C	AGE	INCOME
13276104	22	$14.940	29Jul1998	yes	43	$91.449
8568865	48	$18.510	########	no	40	$106.952
3764824	25	$17.600	02Oct1988	yes	55	$59.162
6916555	7	$17.230	30Jan1995	no	46	$59.162
2693415	12	$25.810	21Jan1991	no	40	$59.162
3969722	38	$16.640	12May1993	yes	39	$35.081
1737378	38	$29.450	27Sep1989	no	43	$145.353
7082153	48	$9.280	02Jan1994	no	52	$0
2589510	22	$6.450	24Nov1998	no	35	$14.508
2556136	23	$29.270	23Nov1993	yes	40	$57.474
7530371	62	$4.600	09Nov1989	yes	31	$26.520
2130575	24	$23.450	22Sep1994	no	42	$52.988
4012002	14	$28.560	12May1986	yes	48	$52.988
9350798	31	$33.710	02Jul1991	no	49	$52.988
2990245	23	$33.320	22Feb1990	yes	40	$52.988
3939254	30	$10.910	########	yes	45	$61.931
5366048	40	$23.230	06Jun1992	yes	48	$61.509
8033252	35	$22.050		no	44	$139.330
4660671	41	$17.470	15Jun1987	no	38	$0
5935828	18	$23.390	07Jul1980	yes	42	$76.226
1704832	36	$17.230	05Jun1997	no	52	$59.162
42131636	8	$30.150	########	no	43	$132.561
3707484	21	$7.500	29Jul1985	yes	40	$125.893
7182942	14	$17.550	21Oct1995	no	40	$123.520
5388757	33	$29.210	02Jun1991	yes	47	$45.257
5209593	53	$13.050	12Dec1993	no	40	$75.516
5684737	40	$36.120	16Dec1990	yes	47	$104.271
4538673	35	$28.180	########	no	33	$111.427
5820861	11	$17.300	########	yes	50	$111.427
6804259	23	$11.620	30Mar1995	no	51	$50.166
93412915	50	$14.530	09Dec1982	no	43	$48.184
46157391	24	$21.990	21Jun1983	no	40	$22.059
2252155	35	$12.180	########	yes	40	$23.571
6833784	45	$27.890	05Nov1989	no	55	$55.409
5039064	48	$8.460	05Sep1987	yes	48	$39.613
19707619	9	$17.230	07Jun1988	no	45	$23.773
3457713	17	$31.390	25Apr1996	yes	40	$55.364
8308556	44	$23.320	23Jan1900	no	55	$163.158
6429873	23	$10.350	21Jan1982	yes	52	$59.162
9309292	45	$27.020	04Apr1992	no	54	$107.808
39176001	42	$7.470	22Dec1987	no	39	$59.685
8419408	51	$6.720	26Jun1991	no	44	$146.267
60189132	10	$6.120	23Apr1986	no	45	$1.158
7921960	32	$13.160	23Oct1984	no	43	$1.158

SIMULATIONS ERGEBNISSE FEHLENDE WERTE

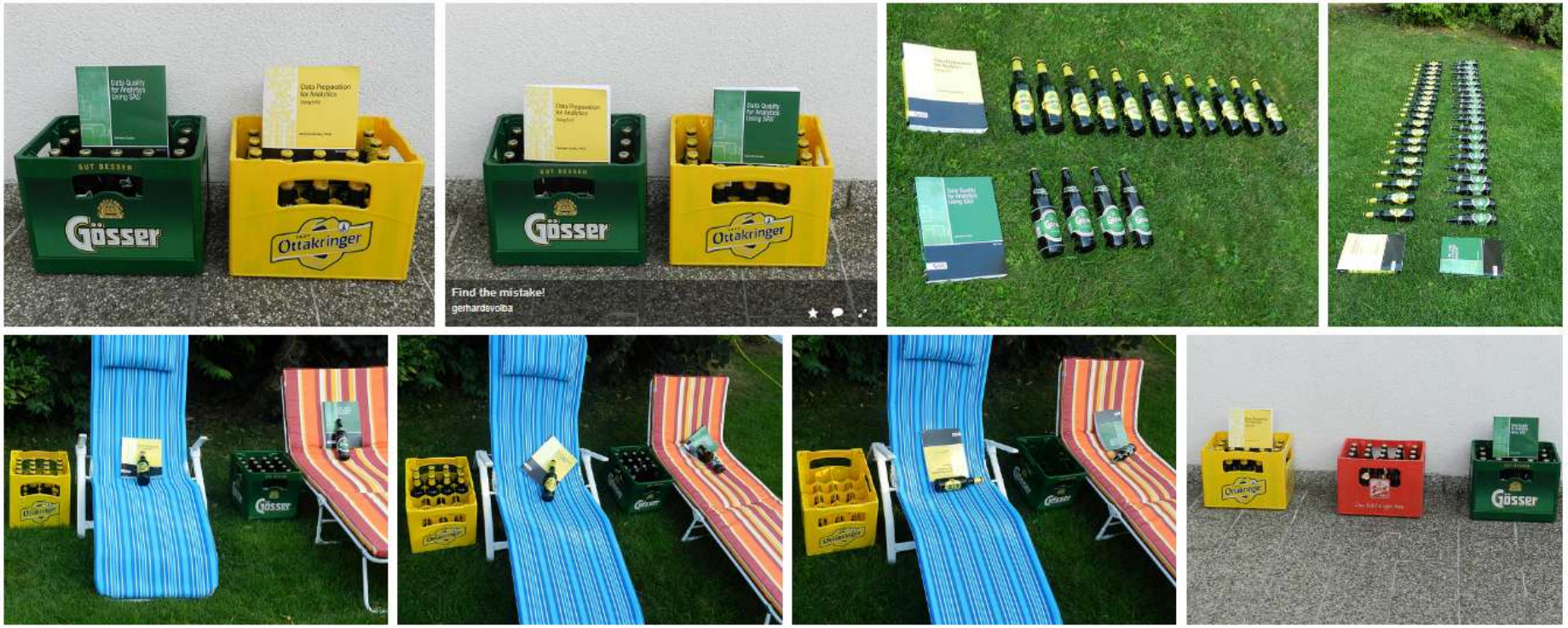
SYSTEMATISCHE MUSTER IN FEHLENDEN WERTEN HABEN EINEN STARKEN EINFLUSS

- Zufällig fehlende Werte nur in den Trainingsdaten haben nur beschränkte Auswirkungen.
- Fehlende Werte in den Scoringdaten beeinflussen das Ergebnis viel stärker.
- Systematische fehlende Werte haben einen viel stärkeren Effekt
- Punkte die wichtig sind:
  - Nicht nur der Anteil fehlender Werte, sondern insbesondere der Typ
  - Fehlende Werte in den Scoring Daten



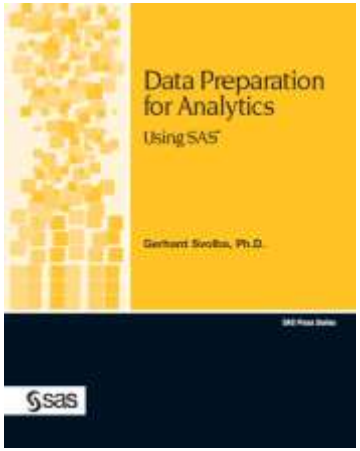
DP AND DQ ON THE ROAD

- http://www.sascommunity.org/wiki/%22Data_Preparation_for_Analytics%22_and_%22Data_Quality_for_Analytics%22_on_the_Road



KONTAKT-INFORMATIONEN

Data Quality for Analytics Using SAS
SAS Press 2012
http://www.sascommunity.org/wiki/Data_Quality_for_Analytics



Data Preparation for Analytics Using SAS
SAS Press 2006
http://www.sascommunity.org/wiki/Data_Preparation_for_Analytics

Gerhard Svolba
Analytic Solution Architect
SAS-Austria
Gerhard.svolba@sas.com
http://www.sascommunity.org/wiki/Gerhard_Svolba
LinkedIn
XING