

22. SAS Club

Wien, 8. November 2012
Gerstenboden der Ottakringer Brauerei



Ottakringer Brauerei, November 2006

14. SAS Club

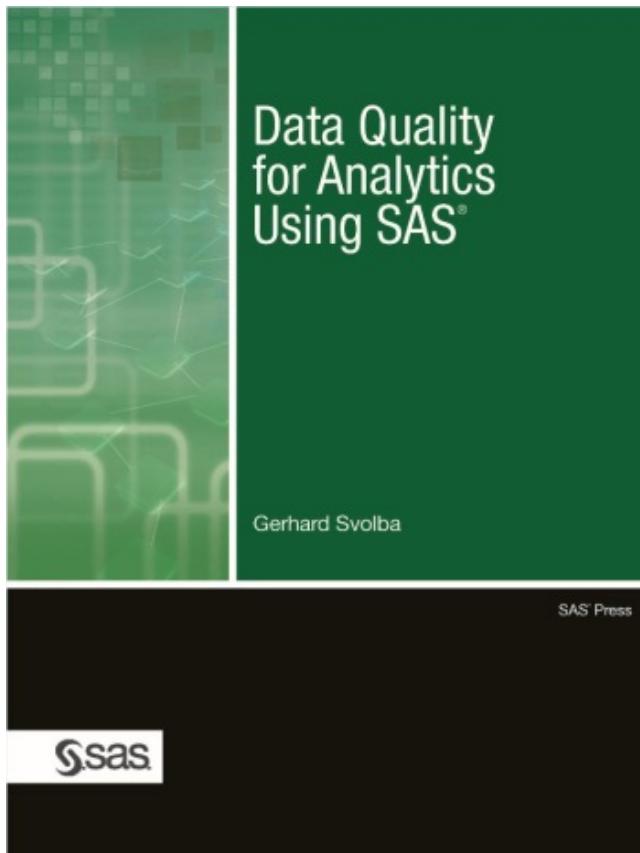


2

Data Quality for Analytics Using SAS

SAS Press, April 2012

Dr. Gerhard Svolba – sastools.by.gerhard@gmx.net – LinkedIn
http://www.sascommunity.org/wiki/Data_Quality_for_Analytics



- Analytik hat zusätzliche Anforderungen an Datenqualität.
- Analytik bietet Methoden für bessere Datenqualität.
- Simulations-Studien zeigen die Konsequenzen von schlechter Datenqualität auf die Modellgenauigkeit.

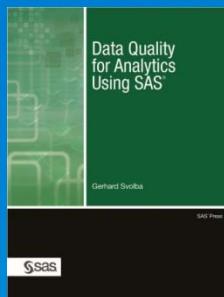
Agenda

- „Datenqualität“ und „Datenqualität für Analytik“ – macht das einen Unterschied?
Dr. Gerhard Svolba
- **SAS Tipps und Tricks:**
Sie wollen rechtzeitig ein Bild über die Datenqualität Ihrer Analysedaten haben? SAS und JMP helfen Ihnen dabei.
Dr. Mihai Paunescu
- Ihre Datenqualität ist nicht so gut, wie sie dachten. Was sind die Konsequenzen für analytische Auswertungen?
Dr. Gerhard Svolba
- **SAS News Corner**
Philipp Manschek
- „Von Bücherverbrennungen, Desserttellern und Farbschablonen“
Dr. Gerhard Svolba

„Datenqualität“ und „Datenqualität für Analytik“ –

Macht das einen Unterschied?

Dr. Gerhard Svolba



„Über den Vortragenden“

- Produktverantwortlicher für die SAS Analytik Produkte
- Analytik Solution Architect 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



www.ycpodersdorf.at

Kleine Segel- und Regattakunde (2)

Aufkreuzen
gegen den Wind.

Mehrmaliges
Wenden, da die
nächste Boje
nicht am direkten
Weg erreicht
werden kann.



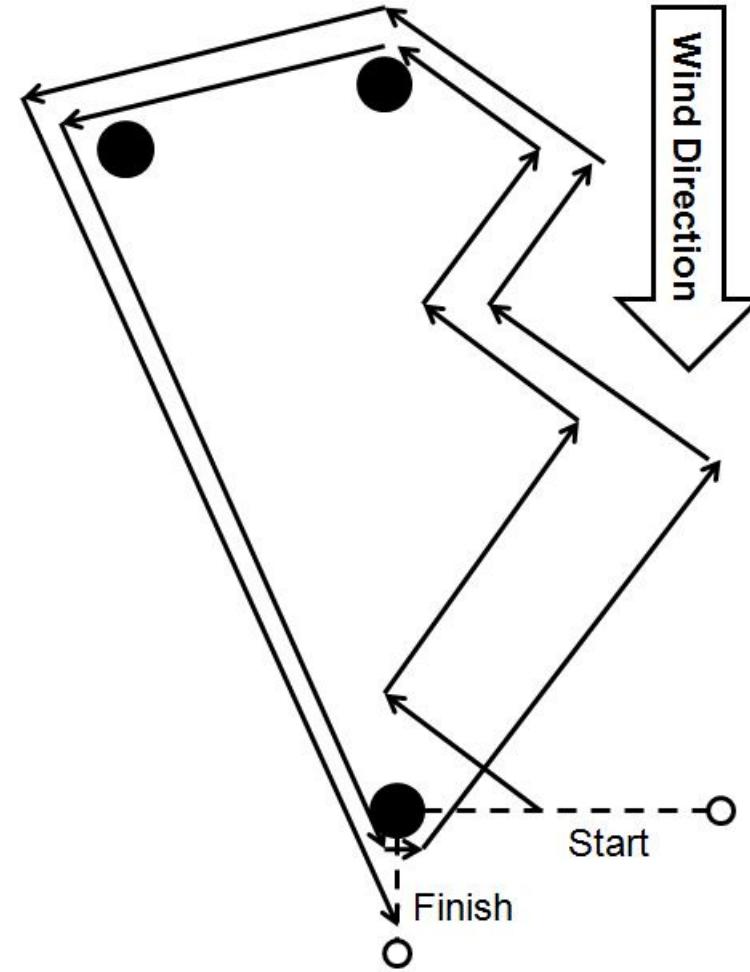
Kleine Segel- und Regattakunde (3)

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



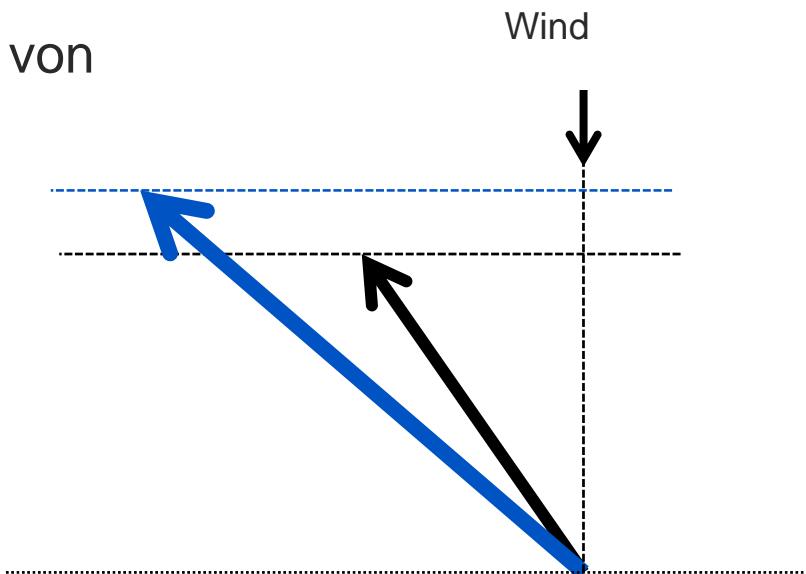
Kleine Segel- und Regattakunde (4)

Skizze einer
Regattabahn mit
3 Bojen.



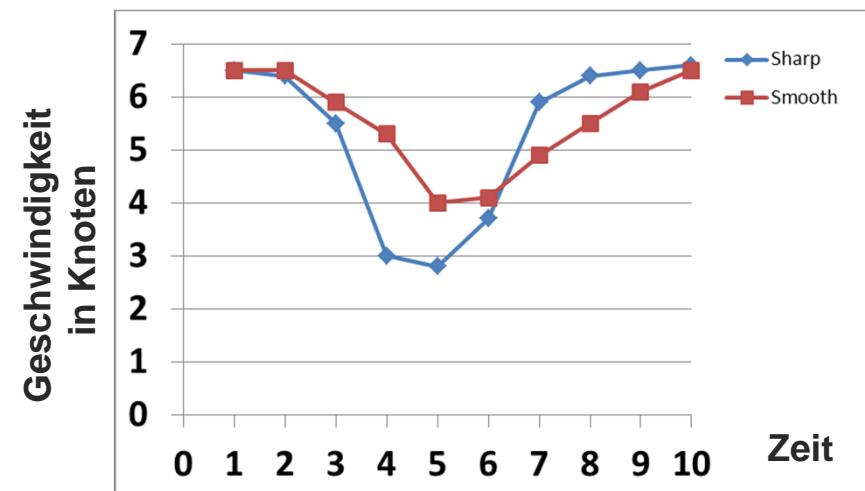
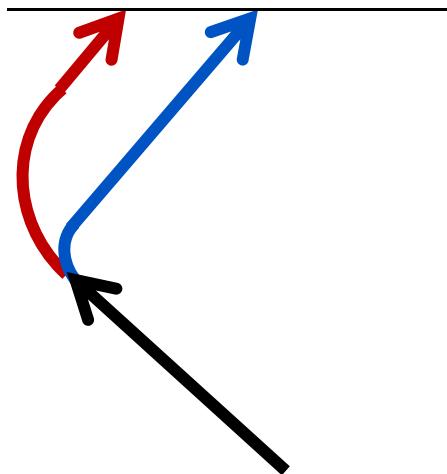
Hier kommt der Statistiker ins Spiel! Analysefragen zur Optimierung der Segeltaktik 1

- 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="www" 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 Wettkunft
- Sonstige Kommentare

Datum	Regatta	Wettkunft	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

14

Datenqualitäts-Probleme in der Case Study und in Business-Analysen sind sich ähnlich

- 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

Datenqualitäts-Probleme in der Case Study und in Business-Analysen sind sich ähnlich

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

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 →
	Luftdruck	1026,5 hPa →
	Relative Luftfeuchte	49 % →
	Niederschlag 1h/24h	0,0 l/m² / 0,2 l/m² →

Quelle: www.byc.at



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

Kategorisierung der Datenqualitätsprobleme (Forts.)

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

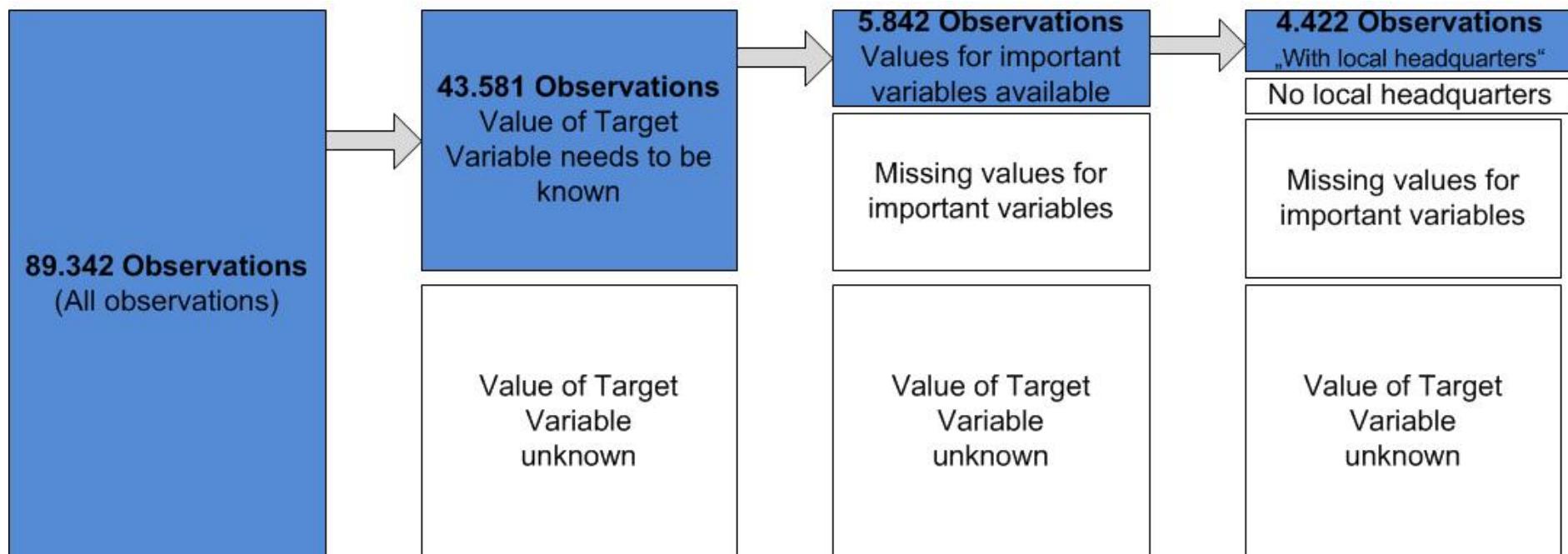
Data
Usability

Data
Availability

Der Begriff „Historische Daten“ muss genau definiert werden

	January					
	22	23	24	25	26	27
Rented Cars	18.912	17.730	17.618	16.708	17.899	16.855
Bookings (per day before)	18.853	17.729	17.616	16.510	17.728	16.843
Bookings (day -2)		17.693	17.617	16.512	17.727	16.881
Bookings (day -3)			17.701	16.511	17.678	16.709
Bookings (day -4)				16.666	17.675	16.707
Bookings (day -5)					17.619	16.513
Bookings (day -6)						16.509

Die Anzahl der für die Analyse verwendbaren Beobachtungen reduziert sich rasch



Der Effekt fehlender Werte auf die Datenquantität

Proportion Missing Values	Number of Variables											
	1	2	3	4	5	10	15	20	25	30	40	50
1%	99.0%	98.0%	97.0%	96.1%	95.1%	90.4%	86.0%	81.8%	77.8%	74.0%	66.9%	60.5%
2%	98.0%	96.0%	94.1%	92.2%	90.4%	81.7%	73.9%	66.8%	60.3%	54.5%	44.6%	36.4%
3%	97.0%	94.1%	91.3%	88.5%	85.9%	73.7%	63.3%	54.4%	46.7%	40.1%	29.6%	21.8%
4%	96.0%	92.2%	88.5%	84.9%	81.5%	66.5%	54.2%	44.2%	36.0%	29.4%	19.5%	13.0%
5%	95.0%	90.3%	85.7%	81.5%	77.4%	59.9%	46.3%	35.8%	27.7%	21.5%	12.9%	7.7%
6%	94.0%	88.4%	83.1%	78.1%	73.4%	53.9%	39.5%	29.0%	21.3%	15.6%	8.4%	4.5%
7%	93.0%	86.5%	80.4%	74.8%	69.6%	48.4%	33.7%	23.4%	16.3%	11.3%	5.5%	2.7%
8%	92.0%	84.6%	77.9%	71.6%	65.9%	43.4%	28.6%	18.9%	12.4%	8.2%	3.6%	1.5%
9%	91.0%	82.8%	75.4%	68.6%	62.4%	38.9%	24.3%	15.2%	9.5%	5.9%	2.3%	0.9%
10%	90.0%	81.0%	72.9%	65.6%	59.0%	34.9%	20.6%	12.2%	7.2%	4.2%	1.5%	0.5%
15%	85.0%	72.3%	61.4%	52.2%	44.4%	19.7%	8.7%	3.9%	1.7%	0.8%	0.2%	0.0%
20%	80.0%	64.0%	51.2%	41.0%	32.8%	10.7%	3.5%	1.2%	0.4%	0.1%	0.0%	0.0%
25%	75.0%	56.2%	42.2%	31.6%	23.7%	5.6%	1.3%	0.3%	0.1%	0.0%	0.0%	0.0%
30%	70.0%	49.0%	34.3%	24.0%	16.8%	2.8%	0.5%	0.1%	0.0%	0.0%	0.0%	0.0%

22

Korrelation zwischen Variablen „Problem und Segen“

- Wichtige Eigenschaft in multivariaten Modellen:
Inputvariablen sollen nicht multikollinear sein.
- Ersetzen von fehlenden Werten und Substituieren
des Effekts nicht vorhandener oder nicht
verwendbarer Variablen

	Domain A (Soziodemographie)			Domain B (Finanzkennzahlen)			Domain C (Produktnutzung)		
Variable	A1	A2	A3	B1	B2	B3	C1	C2	C3
Scenario 1	Value	Value	.	Value	.	Value	Value	.	Value
Scenario 2	Value	Value	Value	.	.	.	Value	Value	Value

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

- Univariante und multivariate Plausibilitätskontrollen

▪ Statistische Eigenschaften

- Korrelation, Variabilität und Verteilung

24

Das sind Ihre Optionen, wenn Sie erkennen, dass die Datenqualität schlecht ist

