The Power to Know.

SAS® USERS GROUP INTERNATIONAL
March 26–29, 2006 | San Francisco

Efficient Construction of a One-Row-per-Subject Data Mart for Data Mining

Gerhard Svolba (PhD)
SAS-Austria (Vienna)

Paper 078-31

Agenda

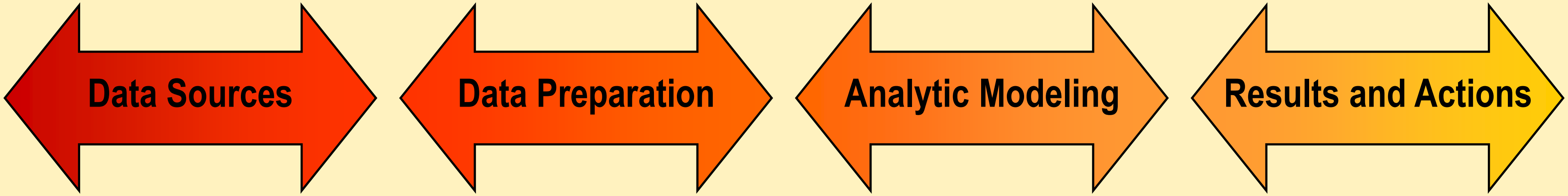
- Analytic Data Preparation
- The One-Row-Per-Subject Paradigm
- Clever Aggregations – Tricky Derived Variables
- Case Study
- Considerations for Predictive Modeling
- Closing Thoughts

Some Words on Analytic Data Preparation

- Is for techies
- Is boring
- Consumes 80 % of the project
- Is something that SAS can excellently do
- Is vital to the quality of the project

- Is presented at 8:00 a.m. after SUGI party

The Analysis Process:
From Raw Data to Actionable Results


Data Sources

Data Preparation

Analytic Modeling

Results and Actions

Different Data Sources

Relational Models, Star Schemes

Merges, Denormalisation

Derived Variables

Transpositions, Aggregations

Modeling, Parameter Estimation, Tuning,

Predictions, Classifications, Clustering

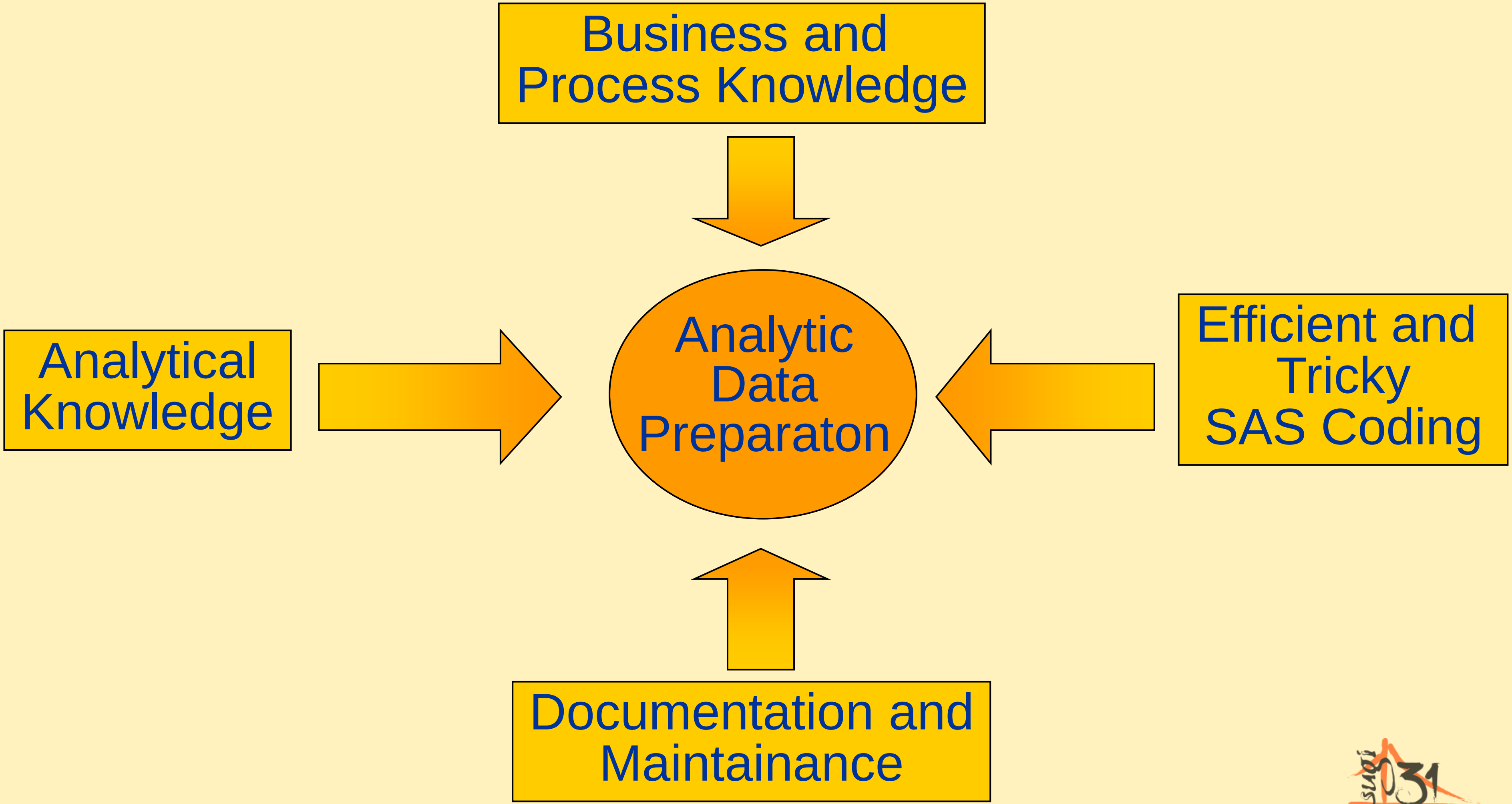
Usage of Results

Profiling

Interpretations

Data Availability + **Adequate Preparation** + Clever Modeling = Good Results

Key Success Factors for Analytic Data Preparation



Business and Process Knowledge

Analytical Knowledge

Analytic Data Preparaton

Efficient and Tricky SAS Coding

Documentation and Maintainance

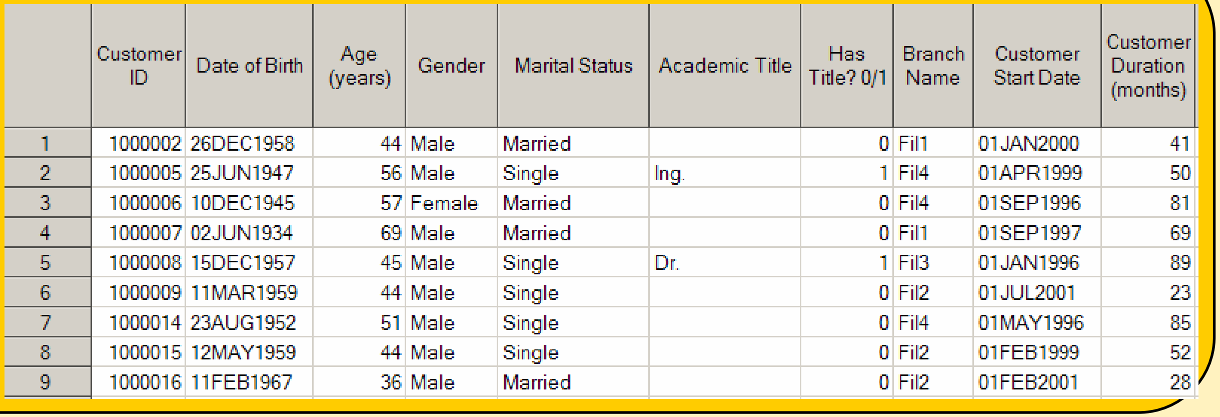
Analysis Subjects and Multiple Observations

- *Analysis subjects* are entities that are being analyzed and the analysis results are interpreted in their context.

- *Multiple observations per analysis subject*
  - Repeated measurements over time
  - Multiple observations because of hierarchical relationships

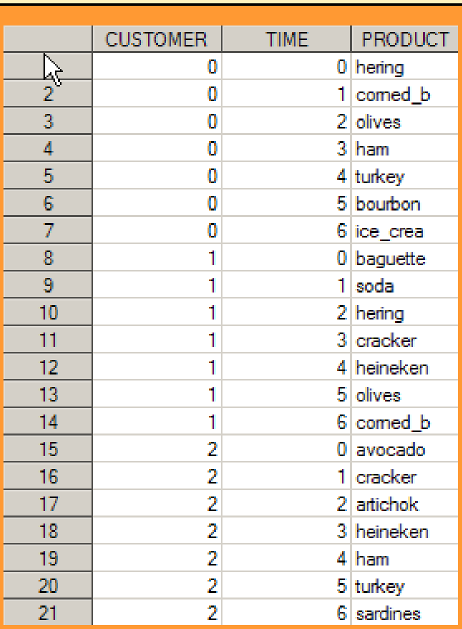
Main Types of Data Marts

One-Row-per-Subject Data Mart



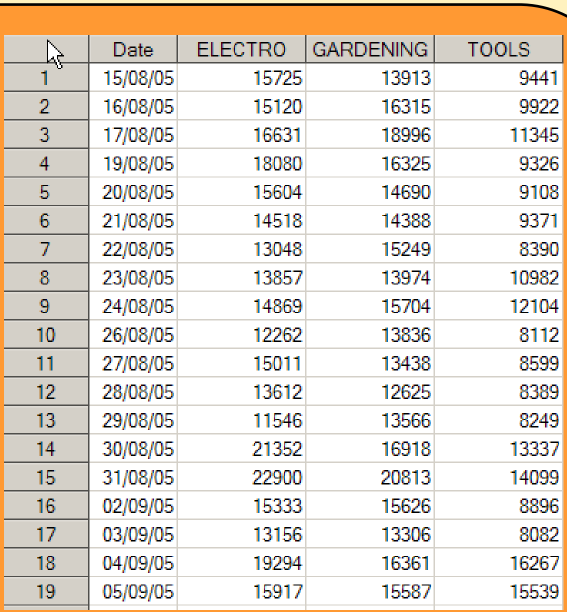
	Customer ID	Date of Birth	Age (years)	Gender	Marital Status	Academic Title	Has Title? 0/1	Branch Name	Customer Start Date	Customer Duration (months)
1	1000002	26DEC1958	44	Male	Married		0	Fil1	01JAN2000	41
2	1000005	25JUN1947	56	Male	Single	Ing.	1	Fil4	01APR1999	50
3	1000006	10DEC1945	57	Female	Married		0	Fil4	01SEP1996	81
4	1000007	02JUN1934	69	Male	Married		0	Fil1	01SEP1997	69
5	1000008	15DEC1957	45	Male	Single	Dr.	1	Fil3	01JAN1996	89
6	1000009	11MAR1959	44	Male	Single		0	Fil2	01JUL2001	23
7	1000014	23AUG1952	51	Male	Single		0	Fil4	01MAY1996	85
8	1000015	12MAY1959	44	Male	Single		0	Fil2	01FEB1999	52
9	1000016	11FEB1967	36	Male	Married		0	Fil2	01FEB2001	28

Multiple-Row-per-Subject Data Mart



	CUSTOMER	TIME	PRODUCT
	0	0	hering
2	0	1	comed_b
3	0	2	olives
4	0	3	ham
5	0	4	turkey
6	0	5	bourbon
7	0	6	ice_crea
8	1	0	baguette
9	1	1	soda
10	1	2	hering
11	1	3	cracker
12	1	4	heineken
13	1	5	olives
14	1	6	comed_b
15	2	0	avocado
16	2	1	cracker
17	2	2	artichok
18	2	3	heineken
19	2	4	ham
20	2	5	turkey
21	2	6	sardines

Longitudinal Data Mart



	Date	ELECTRO	GARDENING	TOOLS
1	15/08/05	15725	13913	9441
2	16/08/05	15120	16315	9922
3	17/08/05	16631	18996	11345
4	19/08/05	18080	16325	9326
5	20/08/05	15604	14690	9108
6	21/08/05	14518	14388	9371
7	22/08/05	13048	15249	8390
8	23/08/05	13857	13974	10982
9	24/08/05	14869	15704	12104
10	26/08/05	12262	13836	8112
11	27/08/05	15011	13438	8599
12	28/08/05	13612	12625	8389
13	29/08/05	11546	13566	8249
14	30/08/05	21352	16918	13337
15	31/08/05	22900	20813	14099
16	02/09/05	15333	15626	8896
17	03/09/05	13156	13306	8082
18	04/09/05	19294	16361	16267
19	05/09/05	15917	15587	15539

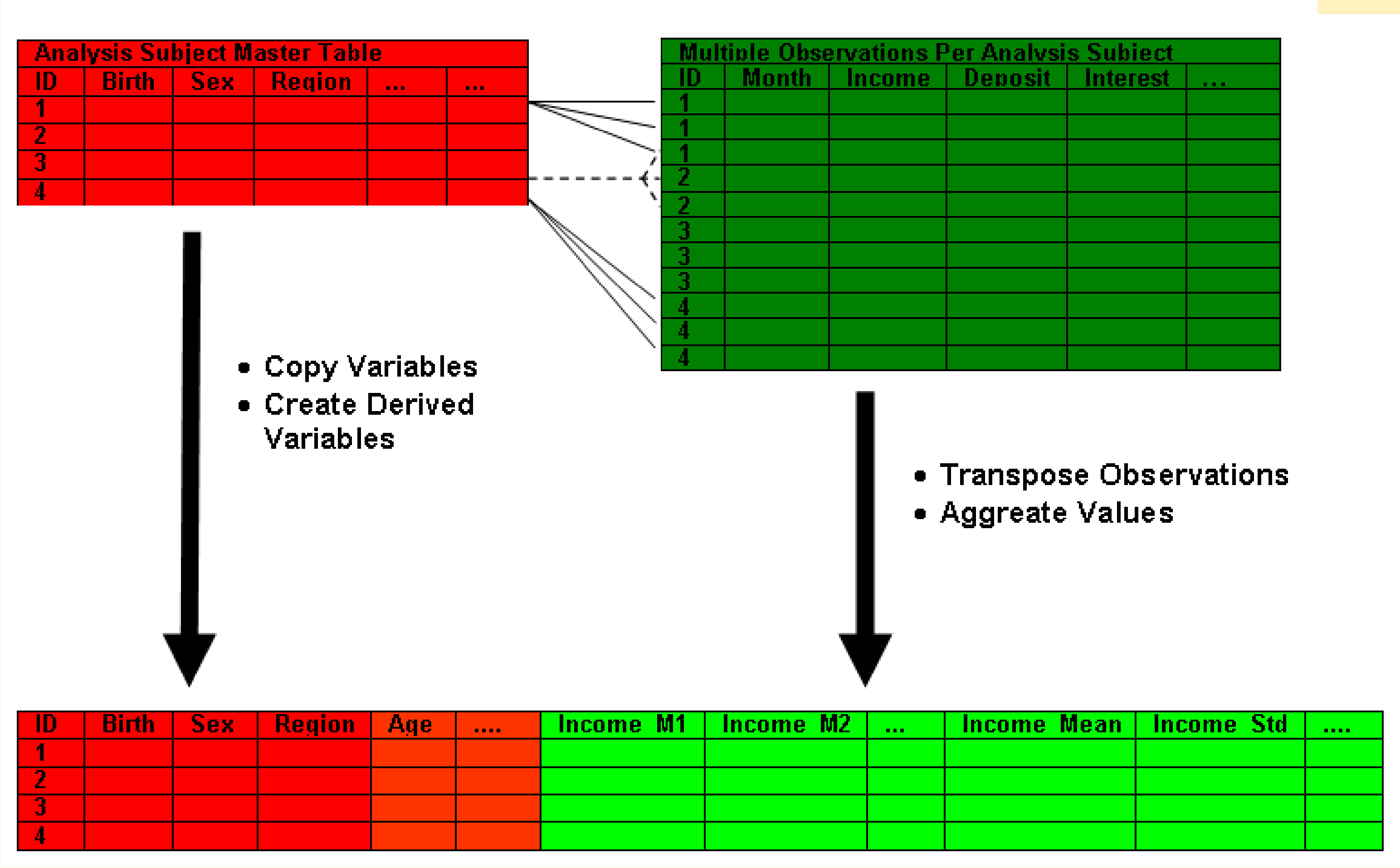
Data Mart Structures

	Data Mart Structure for the Analysis	
Structure of the source data: “Multiple observations per analysis subject exist?”	One-Row-per-Subject Data Mart	Multiple-Row-per-Subject Data Mart
NO		
YES		

The One-Row-Per-Subject Data Mart

- Required by many statistical methods
  - Regression Analysis, Neural Networks, Decision Trees, Survival analysis, Cluster analysis, …
- Most prominent data mart structure in data mining
  - Event prediction (Churn, Fraud, Delinquency, Response, …)
  - Value prediction (Purchase Size, Claim Amount, …)
  - Segmentation (Clustering, …)

The One-Row-Per-Subject Paradigm



Analysis Subject Master Table

ID	Birth	Sex	Region	...	...
1					
2					
3					
4					

Multiple Observations Per Analysis Subject

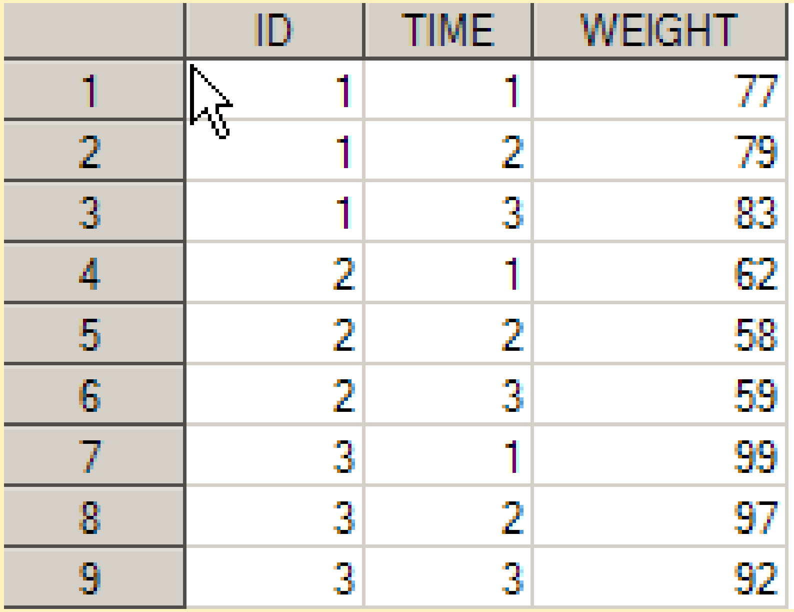
ID	Month	Income	Deposit	Interest	...
1					
1					
1					
2					
2					
3					
3					
3					
4					
4					
4					

- Copy Variables
- Create Derived Variables

- Transpose Observations
- Aggregate Values

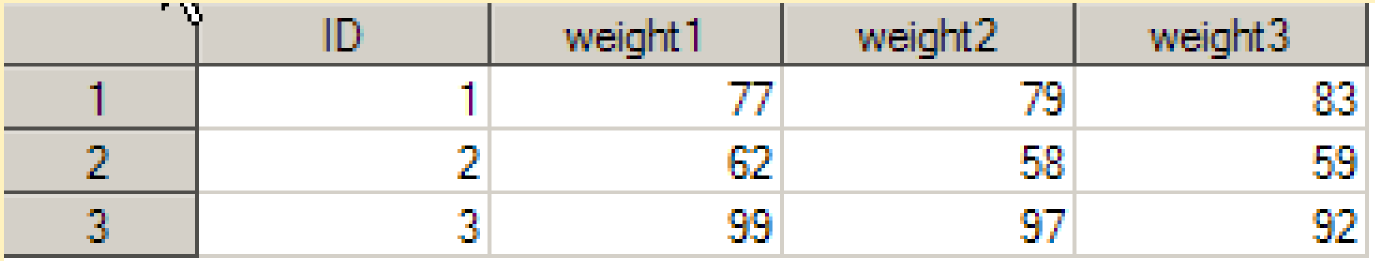
ID	Birth	Sex	Region	Age	....	Income M1	Income M2	...	Income Mean	Income Std	....
1											
2											
3											
4											

Transposing Data to One-Row-Per-Subject



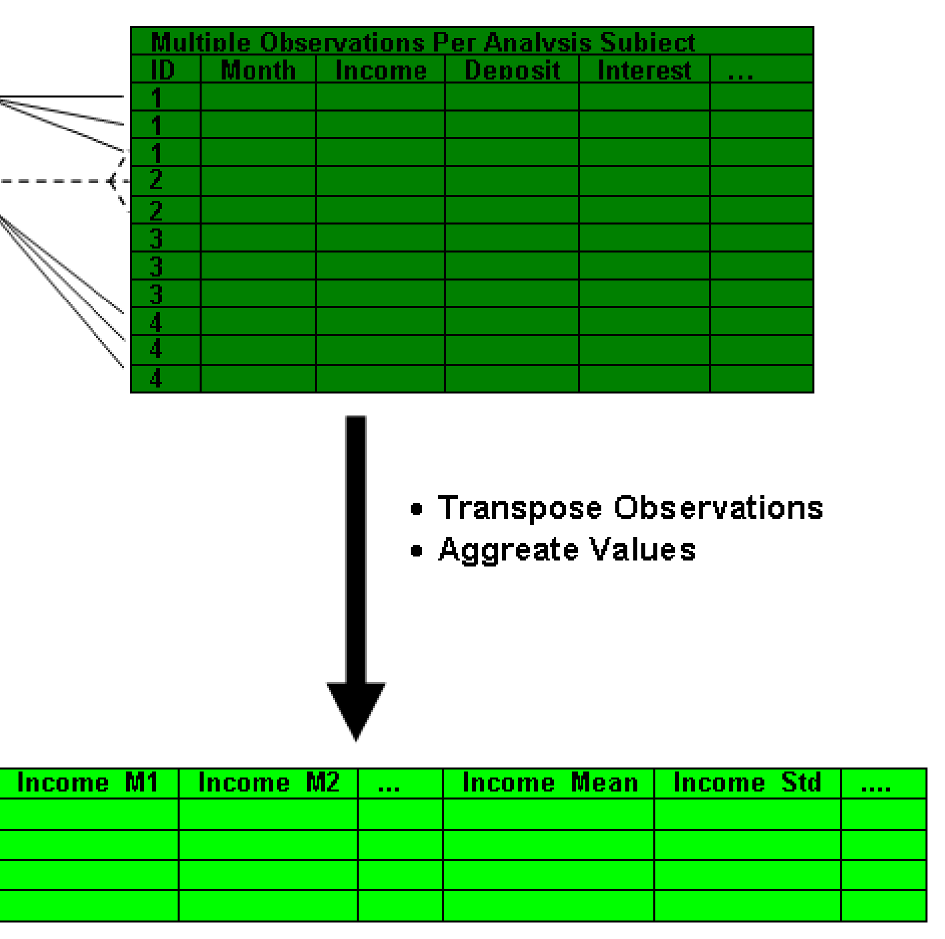
	ID	TIME	WEIGHT
1	1	1	77
2	1	2	79
3	1	3	83
4	2	1	62
5	2	2	58
6	2	3	59
7	3	1	99
8	3	2	97
9	3	3	92

```
PROC TRANSPOSE DATA = long
               OUT = wide(DROP = _name_)
               PREFIX = weight;
 BY id ;
 VAR weight;
 ID time;
RUN;
```



	ID	weight1	weight2	weight3
1	1	77	79	83
2	2	62	58	59
3	3	99	97	92

Clever Aggregations



Multiple Observations Per Analysis Subject					
ID	Month	Income	Deposit	Interest	...
1					
1					
1					
2					
2					
3					
3					
3					
4					
4					
4					

- Transpose Observations
- Aggregate Values

Income M1	Income M2	...	Income Mean	Income Std	....

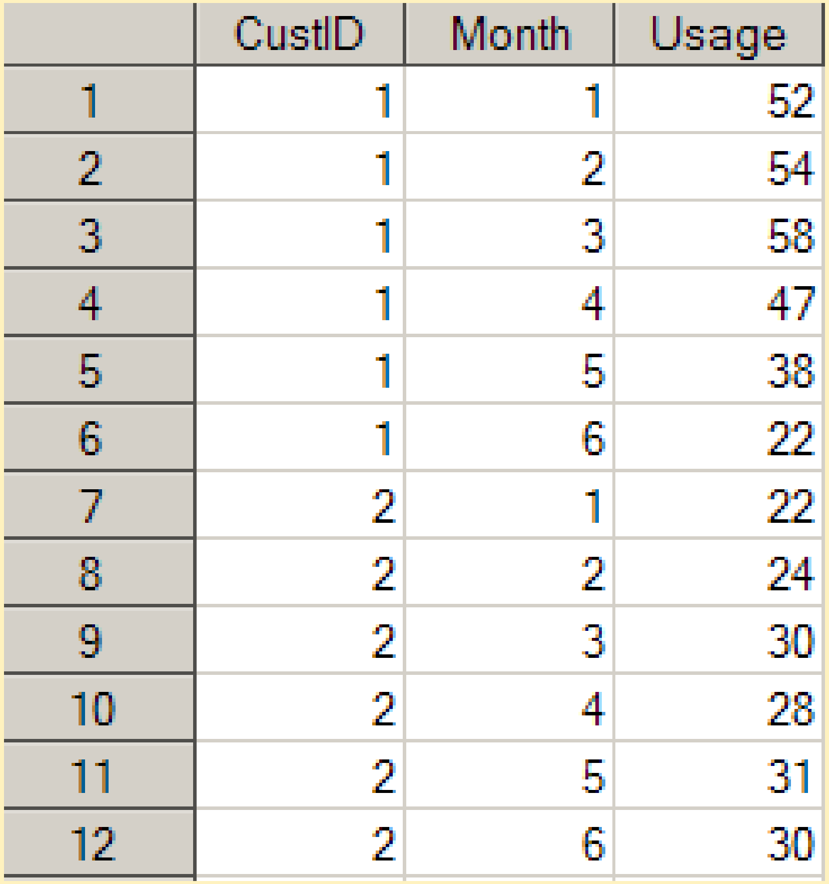
Interval Data

- Static Aggregation
- Correlation of Values
- Course over Time
- Concentration of Values

Categorical Data

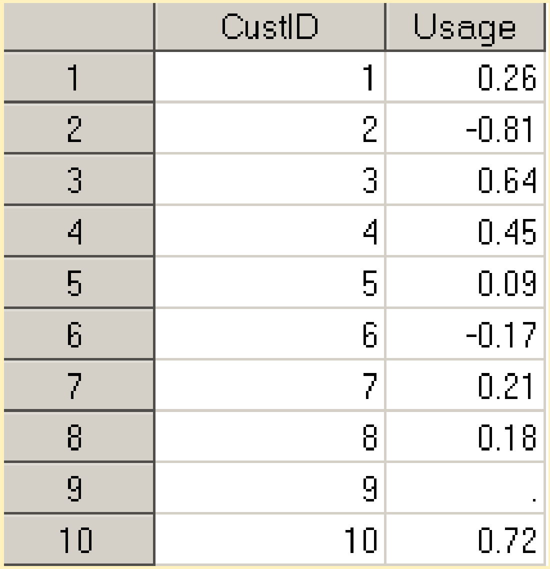
- Frequency Counts
- Concatenated Frequencies
- Total and Distinct Counts

Correlation of Values



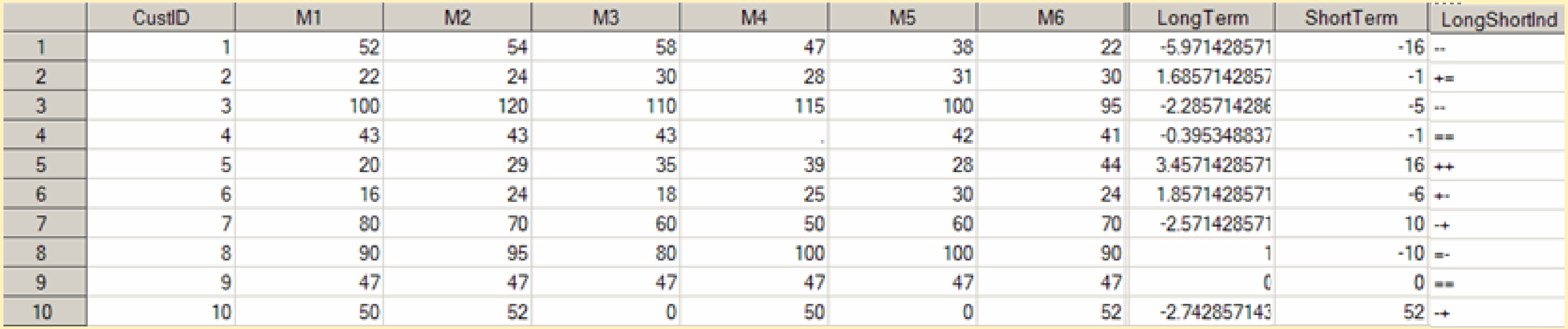
	CustID	Month	Usage
1	1	1	52
2	1	2	54
3	1	3	58
4	1	4	47
5	1	5	38
6	1	6	22
7	2	1	22
8	2	2	24
9	2	3	30
10	2	4	28
11	2	5	31
12	2	6	30

How do the monthly values per subject correlate with the overall mean per month?



	CustID	Usage
1	1	0.26
2	2	-0.81
3	3	0.64
4	4	0.45
5	5	0.09
6	6	-0.17
7	7	0.21
8	8	0.18
9	9	.
10	10	0.72

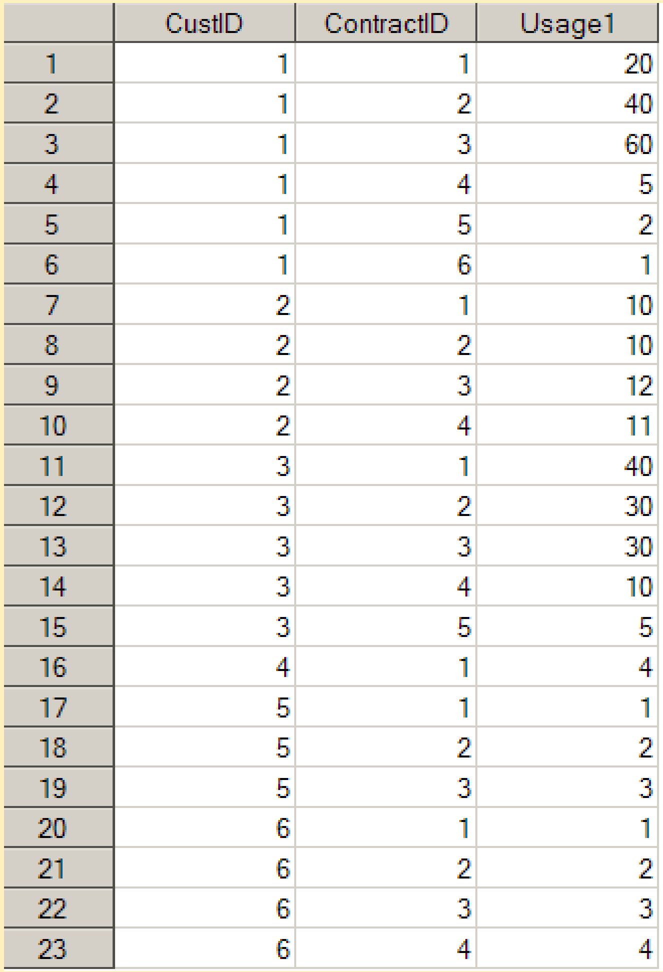
Measures for the Course over Time



	CustID	M1	M2	M3	M4	M5	M6	LongTerm	ShortTerm	LongShortInd
1	1	52	54	58	47	38	22	-5.971428571	-16	--
2	2	22	24	30	28	31	30	1.6857142857	-1	+=
3	3	100	120	110	115	100	95	-2.285714286	-5	--
4	4	43	43	43	.	42	41	-0.395348837	-1	==
5	5	20	29	35	39	28	44	3.4571428571	16	++
6	6	16	24	18	25	30	24	1.8571428571	-6	+-
7	7	80	70	60	50	60	70	-2.571428571	10	-+
8	8	90	95	80	100	100	90	1	-10	=-
9	9	47	47	47	47	47	47	0	0	==
10	10	50	52	0	50	0	52	-2.742857143	52	-+

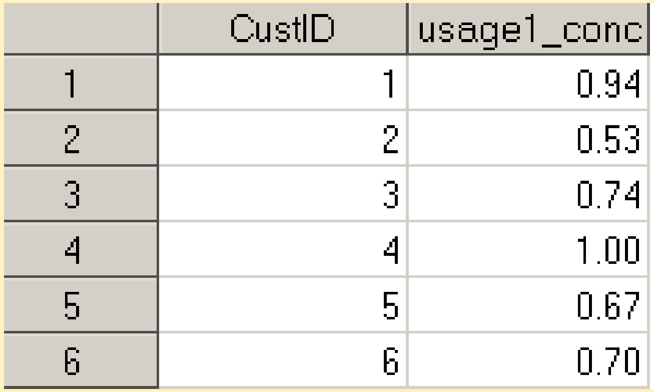
```
PROC REG DATA = longitud NOPRINT
         OUTEST=Est_LongTerm(KEEP = CustID month
                             RENAME = (month=LongTerm));
 MODEL usage = month;
 BY CustID;
RUN;
PROC REG DATA = longitud NOPRINT
         OUTEST=Est_ShortTerm(KEEP = CustID month
                              RENAME = (month=ShortTerm));
 MODEL usage = month;
 BY CustID;
 WHERE month in (5 6);
RUN;
```

Concentration of Values



	CustID	ContractID	Usage1
1	1	1	20
2	1	2	40
3	1	3	60
4	1	4	5
5	1	5	2
6	1	6	1
7	2	1	10
8	2	2	10
9	2	3	12
10	2	4	11
11	3	1	40
12	3	2	30
13	3	3	30
14	3	4	10
15	3	5	5
16	4	1	4
17	5	1	1
18	5	2	2
19	5	3	3
20	6	1	1
21	6	2	2
22	6	3	3
23	6	4	4

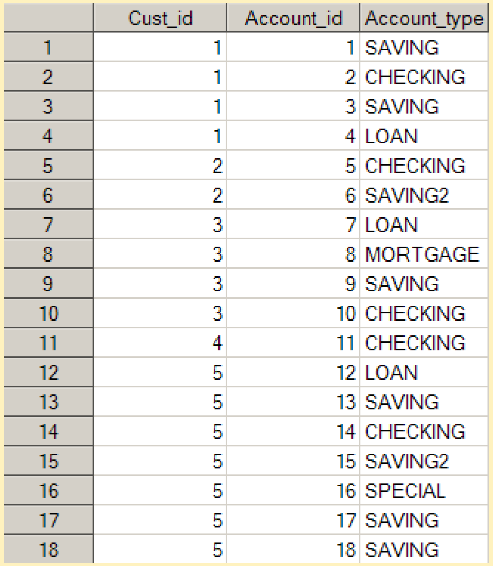
Concentration =
proportion of the sum of the top 50 % sub-hierarchies
/
the total sum over all sub hierarchies



	CustID	usage1_conc
1	1	0.94
2	2	0.53
3	3	0.74
4	4	1.00
5	5	0.67
6	6	0.70

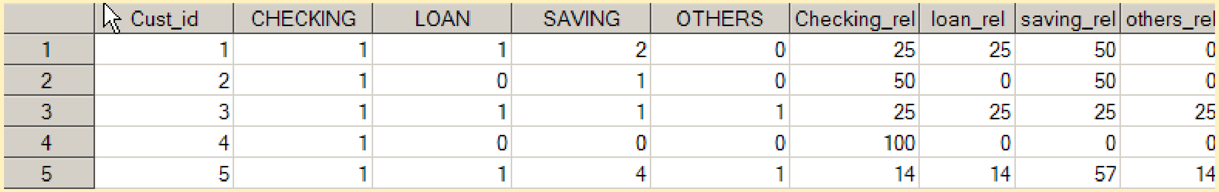
Categorical Variables: Frequency Counts

Source Data



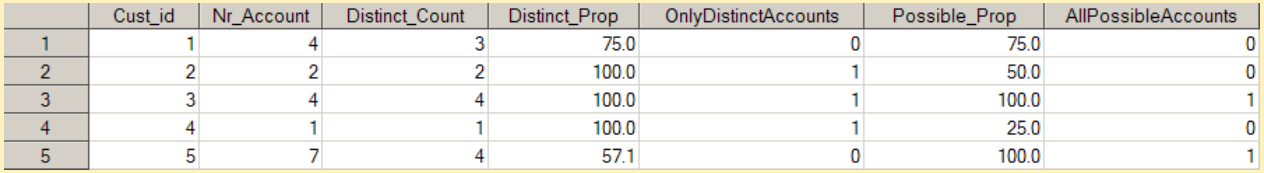
	Cust_id	Account_id	Account_type
1	1	1	SAVING
2	1	2	CHECKING
3	1	3	SAVING
4	1	4	LOAN
5	2	5	CHECKING
6	2	6	SAVING2
7	3	7	LOAN
8	3	8	MORTGAGE
9	3	9	SAVING
10	3	10	CHECKING
11	4	11	CHECKING
12	5	12	LOAN
13	5	13	SAVING
14	5	14	CHECKING
15	5	15	SAVING2
16	5	16	SPECIAL
17	5	17	SAVING
18	5	18	SAVING

Absolute and Relative Frequencies



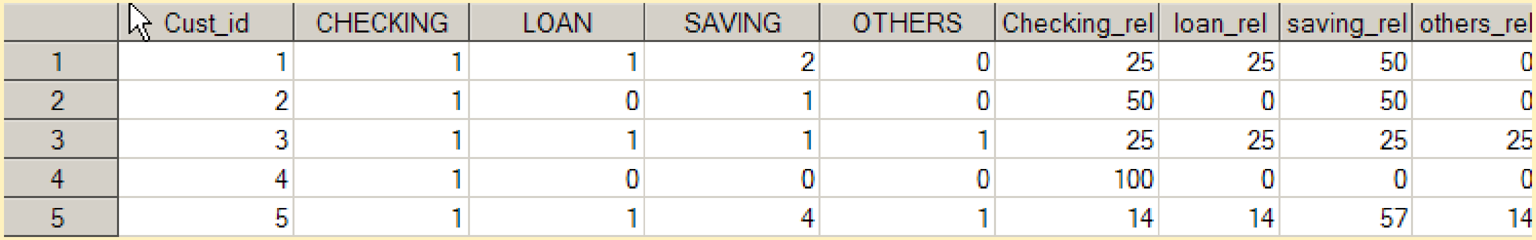
	Cust_id	CHECKING	LOAN	SAVING	OTHERS	Checking_rel	loan_rel	saving_rel	others_rel
1	1	1	1	2	0	25	25	50	0
2	2	1	0	1	0	50	0	50	0
3	3	1	1	1	1	25	25	25	25
4	4	1	0	0	0	100	0	0	0
5	5	1	1	4	1	14	14	57	14

Counts and Distinct Counts



	Cust_id	Nr_Account	Distinct_Count	Distinct_Prop	OnlyDistinctAccounts	Possible_Prop	AllPossibleAccounts
1	1	4	3	75.0	0	75.0	0
2	2	2	2	100.0	1	50.0	0
3	3	4	4	100.0	1	100.0	1
4	4	1	1	100.0	1	25.0	0
5	5	7	4	57.1	0	100.0	1

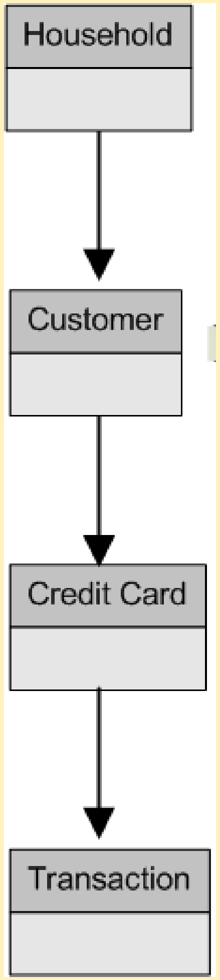
Categorical Variables: Concatenated Frequencies



	Cust_id	CHECKING	LOAN	SAVING	OTHERS	Checking_rel	loan_rel	saving_rel	others_rel
1	1	1	1	2	0	25	25	50	0
2	2	1	0	1	0	50	0	50	0
3	3	1	1	1	1	25	25	25	25
4	4	1	0	0	0	100	0	0	0
5	5	1	1	4	1	14	14	57	14

```
Account_                                    Cumulative    Cumulative
RowPct          Frequency     Percent        Frequency       Percent
--------------------------------------------------------------------
 0_100_0_0          12832       30.61            12832         30.61
 100_0_0_0           9509       22.69            22341         53.30
 50_0_0_50           4898       11.69            27239         64.98
 33_0_0_67           1772        4.23            29011         69.21
 0_0_100_0           1684        4.02            30695         73.23
 67_0_0_33           1426        3.40            32121         76.63
 0_0_50_50            861        2.05            32982         78.69
 50_0_50_0            681        1.62            33663         80.31
```

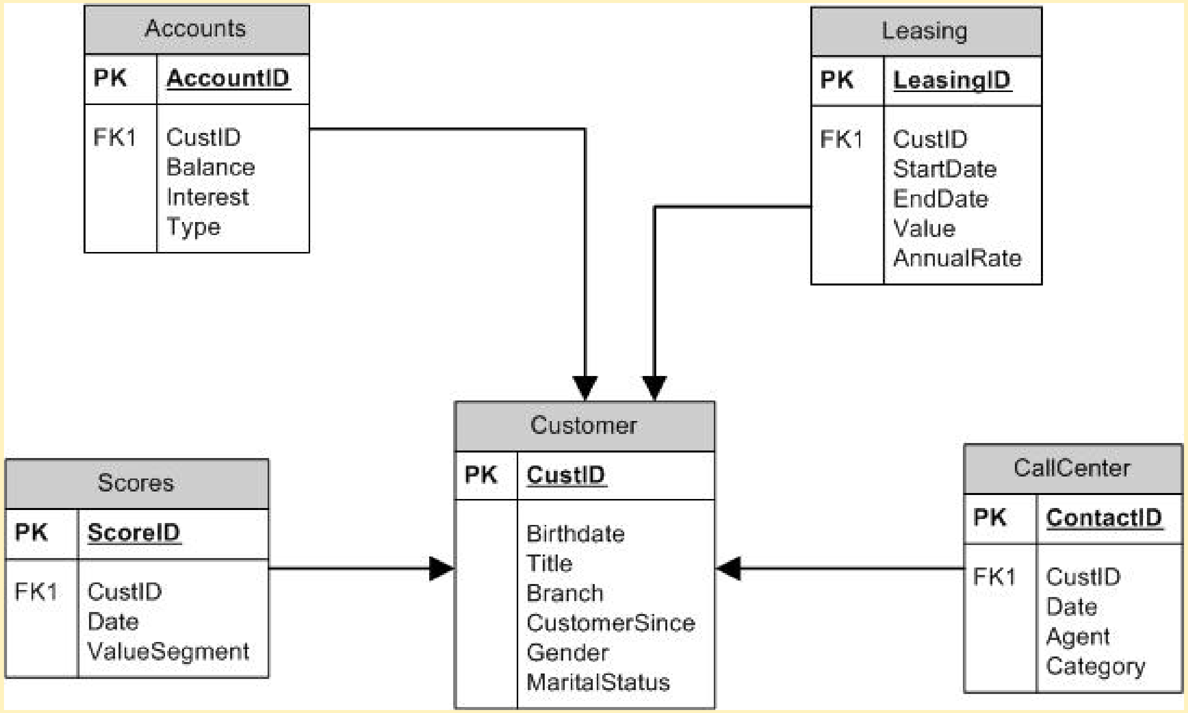
Hierarchies: Aggregating Up, Copying Down



Case Study: Business Question

- Predict customers that have a high probability to leave the company
- Derive target variable „Cancellation YES/NO" from the monthly value segment history (Entry „8. LOST")
- Create a one-row-per-subject data mart for data mining analysis

Case Study: Data and Data Model



Accounts

PK	AccountID
FK1	CustID Balance Interest Type

Leasing

PK	LeasingID
FK1	CustID StartDate EndDate Value AnnualRate

Customer

PK	CustID
	Birthdate Title Branch CustomerSince Gender MaritalStatus

Scores

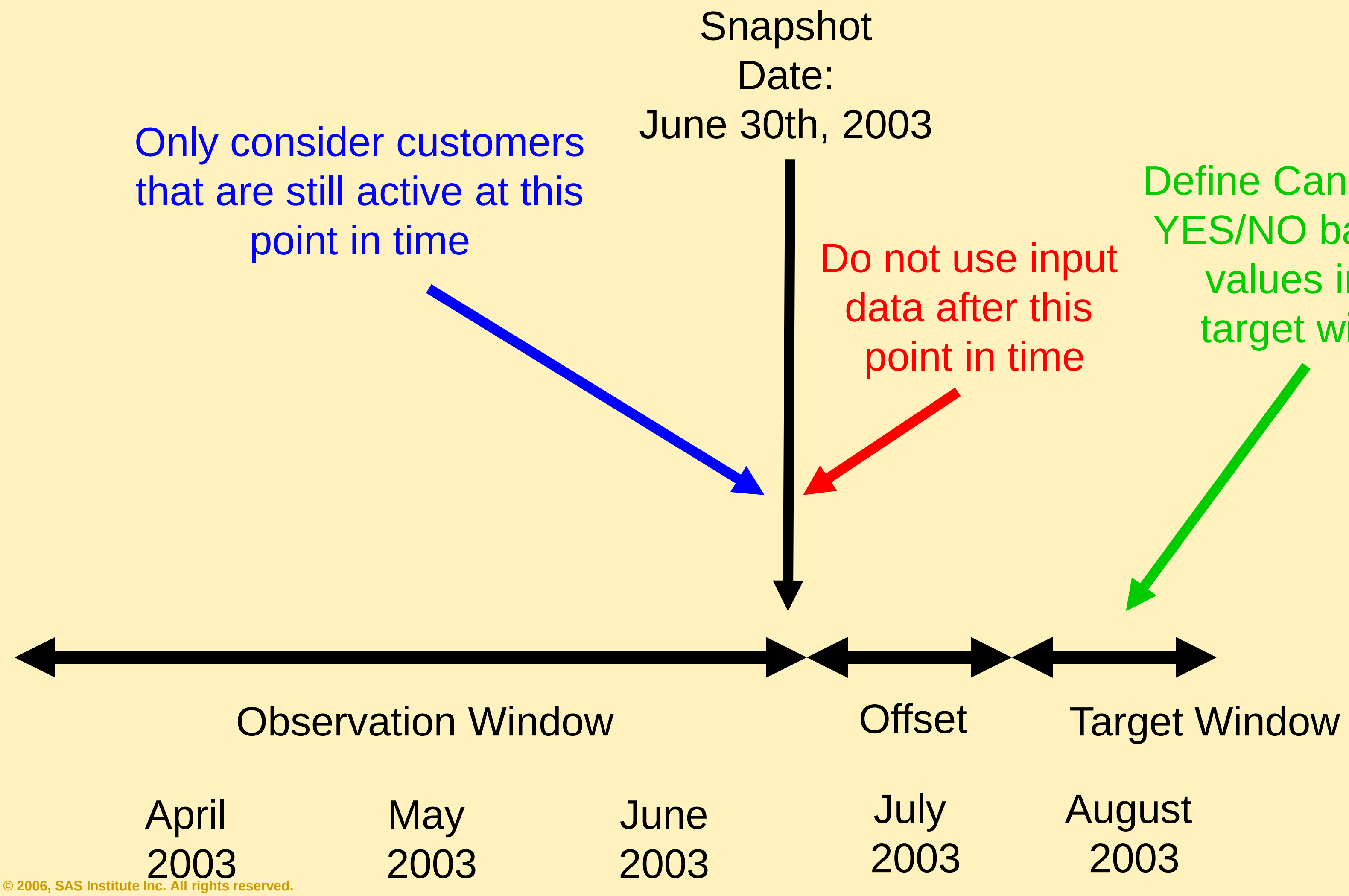
PK	ScoreID
FK1	CustID Date ValueSegment

CallCenter

PK	ContactID
FK1	CustID Date Agent Category

- *Customer data*: demographic and customer baseline data
- *Account data*: information customer accounts
- *Leasing data*: data on leasing information
- *Call Center data*: data on Call center contacts
- *Score data*: data of value segment scores

Considerations for Predictive Modeling



Snapshot
Date:
June 30th, 2003

Only consider customers that are still active at this point in time

Do not use input data after this point in time

Define Cancellation YES/NO based on values in the target window

Observation Window

Offset

Target Window

April
2003

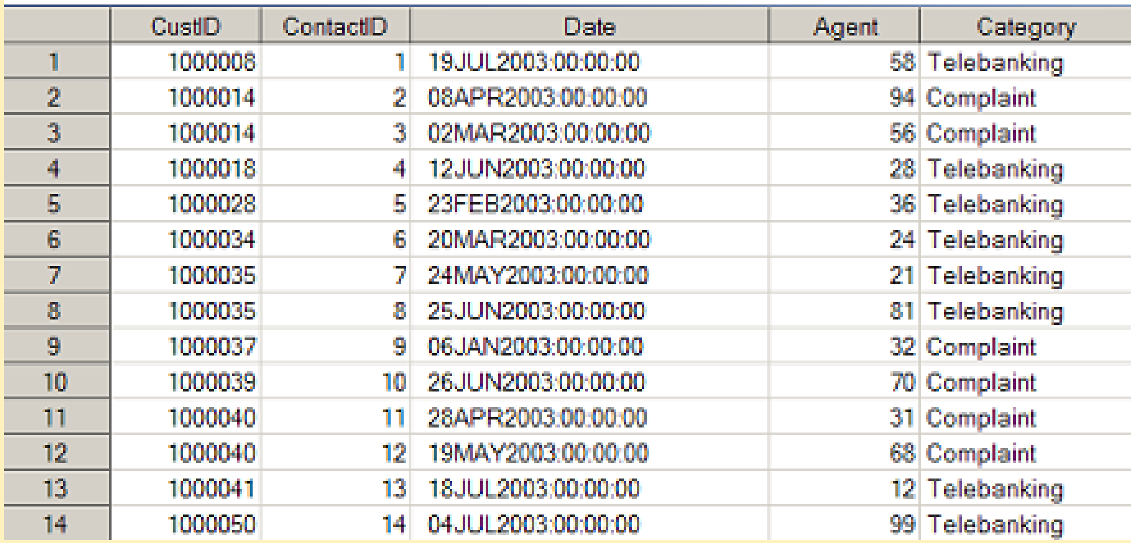
May
2003

June
2003

July
2003

August
2003

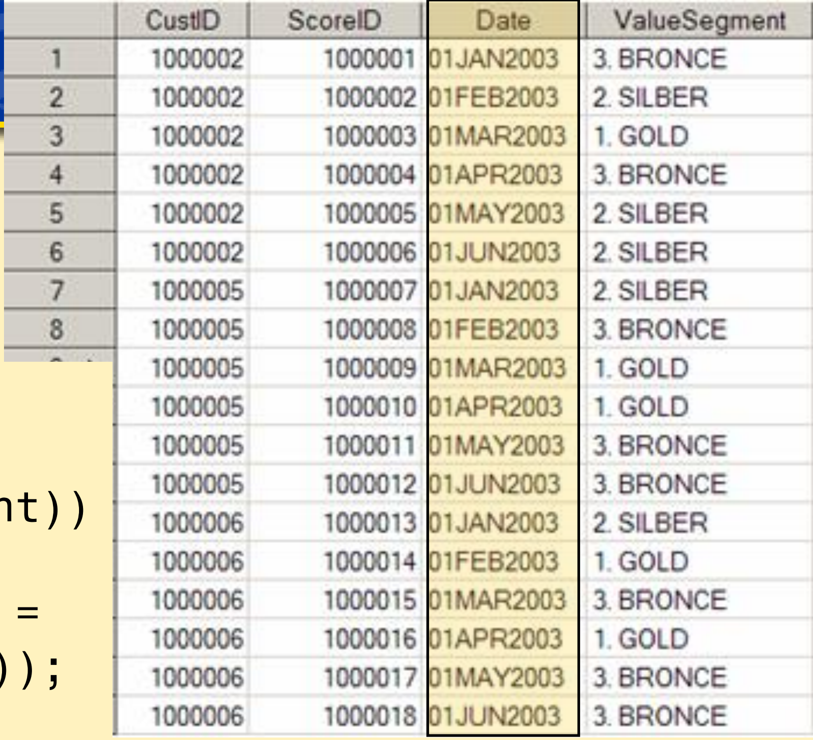
Using Data from the CALLCENTER Table



	CustID	ContactID	Date	Agent	Category
1	1000008	1	19JUL2003:00:00:00	58	Telebanking
2	1000014	2	08APR2003:00:00:00	94	Complaint
3	1000014	3	02MAR2003:00:00:00	56	Complaint
4	1000018	4	12JUN2003:00:00:00	28	Telebanking
5	1000028	5	23FEB2003:00:00:00	36	Telebanking
6	1000034	6	20MAR2003:00:00:00	24	Telebanking
7	1000035	7	24MAY2003:00:00:00	21	Telebanking
8	1000035	8	25JUN2003:00:00:00	81	Telebanking
9	1000037	9	06JAN2003:00:00:00	32	Complaint
10	1000039	10	26JUN2003:00:00:00	70	Complaint
11	1000040	11	28APR2003:00:00:00	31	Complaint
12	1000040	12	19MAY2003:00:00:00	68	Complaint
13	1000041	13	18JUL2003:00:00:00	12	Telebanking
14	1000050	14	04JUL2003:00:00:00	99	Telebanking

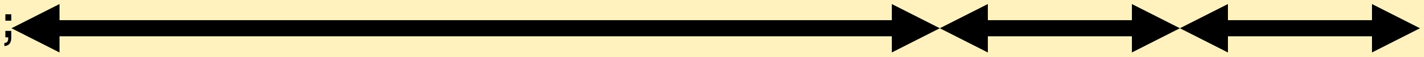
```
%let snapdate = ’30JUN2003’d;
PROC FREQ DATA = callcenter NOPRINT;
 TABLE CustID / OUT = CallCenterComplaints
               (DROP = Percent RENAME =
                     (Count = Complaints));
 WHERE Category = 'Complaint' and
       datepart(date) <= &snapdate;
RUN;
```

Using Data from the SCORES-Table



	CustID	ScoreID	Date	ValueSegment
1	1000002	1000001	01JAN2003	3. BRONCE
2	1000002	1000002	01FEB2003	2. SILBER
3	1000002	1000003	01MAR2003	1. GOLD
4	1000002	1000004	01APR2003	3. BRONCE
5	1000002	1000005	01MAY2003	2. SILBER
6	1000002	1000006	01JUN2003	2. SILBER
7	1000005	1000007	01JAN2003	2. SILBER
8	1000005	1000008	01FEB2003	3. BRONCE
	1000005	1000009	01MAR2003	1. GOLD
	1000005	1000010	01APR2003	1. GOLD
	1000005	1000011	01MAY2003	3. BRONCE
	1000005	1000012	01JUN2003	3. BRONCE
	1000006	1000013	01JAN2003	2. SILBER
	1000006	1000014	01FEB2003	1. GOLD
	1000006	1000015	01MAR2003	3. BRONCE
	1000006	1000016	01APR2003	1. GOLD
	1000006	1000017	01MAY2003	3. BRONCE
	1000006	1000018	01JUN2003	3. BRONCE

```
%let snapdate = '30JUN2003'd;
DATA ScoreFuture(RENAME = (ValueSegment =
                          FutureValueSegment))
     ScoreActual
     ScoreLastMonth(RENAME = (ValueSegment =
                          LastValueSegment));
 SET Scores;
 DATE = INTNX('MONTH',Date,0,'END');
 DROP Date;
 IF Date = &snapdate THEN OUTPUT ScoreActual;
 ELSE IF Date = INTNX('MONTH',&snapdate,-1)
               THEN OUTPUT ScoreLastMonth;
 ELSE IF Date = INTNX('MONTH',&snapdate,2)
               THEN OUTPUT ScoreFuture;
RUN;
```



Observation Window

Offset

Target Window

April 2003

May 2003

June 2003

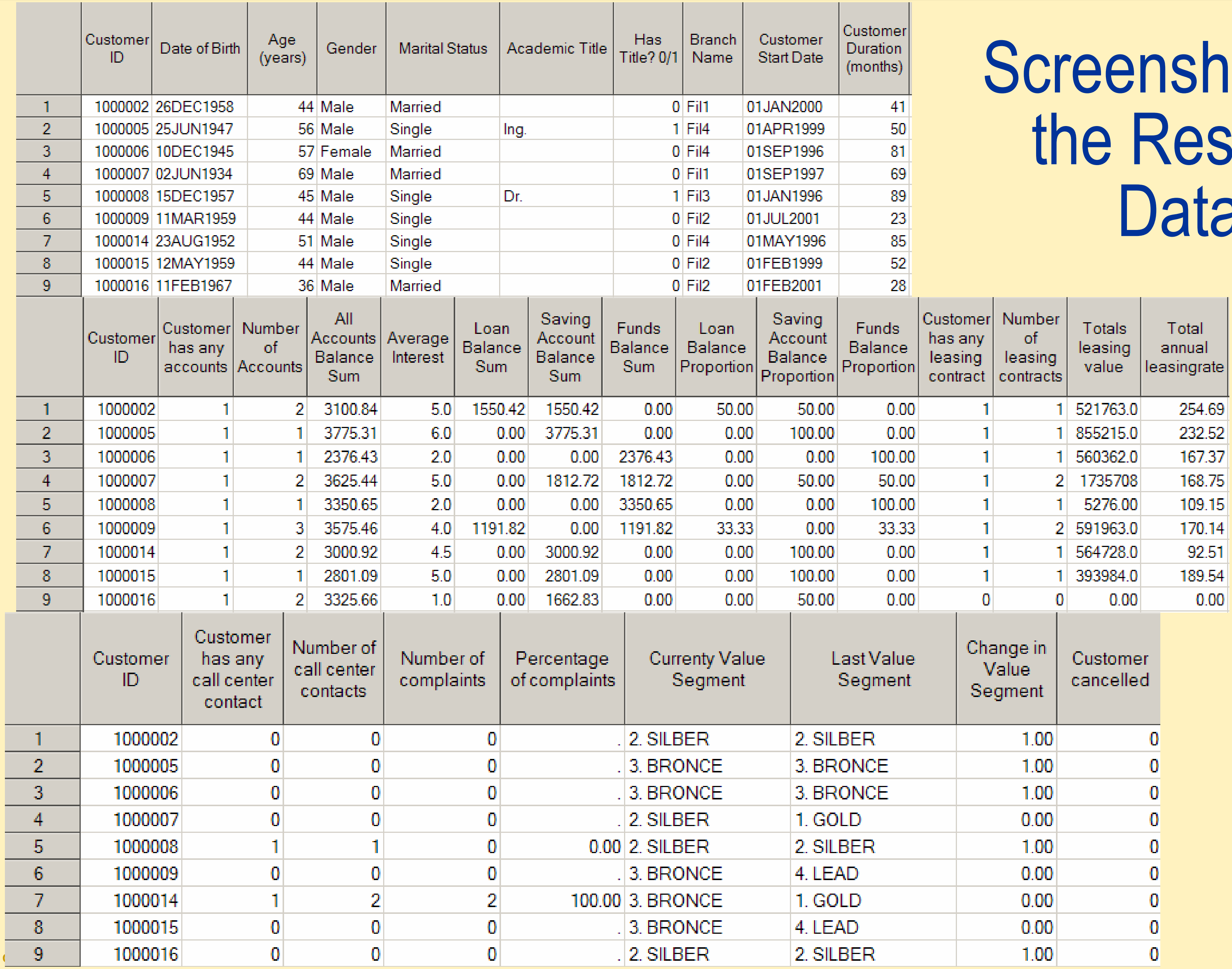
July 2003

August 2003

```
DATA CustomerMart;
ATTRIB /* Customer Baseline */
CustID         FORMAT  = 8.     LABEL = "Customer ID"
Birthdate      FORMAT = DATE9. LABEL = "Date of Birth"
Alter          FORMAT = 8.     LABEL = "Age (years)"
Gender         FORMAT = $6.    LABEL = "Gender"
MaritalStatus  FORMAT = $10.   LABEL = "Marital Status"
Title          FORMAT = $10.   LABEL = "Academic Title"
HasTitle       FORMAT = 8.     LABEL = "Has Title? 0/1"
Branch         FORMAT = $5.    LABEL = "Branch Name";
MERGE Customer (IN = InCustomer)
      AccountSum (IN = InAccounts)
      AccountTypes
      LeasingSum (IN = InLeasing)
      CallCenterContacts (IN = InCallCenter)
      CallCenterComplaints
      ScoreFuture
      ScoreActual
      ScoreLastMonth;
 BY CustID;
 IF InCustomer;
```

```
/* Customer Baseline */
HasTitle = (Title ne " ");
Alter = (&Snapdate-Birthdate)/365.25;
CustomerMonths = (&Snapdate- CustomerSince)/(365.25/12);
/* Accounts */
HasAccounts = InAccounts;
LoanPct = Loan / BalanceSum * 100;
SavingAccountPct = SavingAccount / BalanceSum * 100;
FundsPct = Funds / BalanceSum * 100;
/* Leasing */
HasLeasing = InLeasing;
/* Call Center */
HasCallCenter = InCallCenter;
ComplaintPct = Complaints / Calls *100;
/* Value Segment */
Cancel = (FutureValueSegment = '8. LOST');
ChangeValueSegment = (ValueSegment = LastValueSegment);
RUN;
```

Screenshots of the Resulting Data Mart



	Customer ID	Date of Birth	Age (years)	Gender	Marital Status	Academic Title	Has Title? 0/1	Branch Name	Customer Start Date	Customer Duration (months)
1	1000002	26DEC1958	44	Male	Married		0	Fil1	01JAN2000	41
2	1000005	25JUN1947	56	Male	Single	Ing.	1	Fil4	01APR1999	50
3	1000006	10DEC1945	57	Female	Married		0	Fil4	01SEP1996	81
4	1000007	02JUN1934	69	Male	Married		0	Fil1	01SEP1997	69
5	1000008	15DEC1957	45	Male	Single	Dr.	1	Fil3	01JAN1996	89
6	1000009	11MAR1959	44	Male	Single		0	Fil2	01JUL2001	23
7	1000014	23AUG1952	51	Male	Single		0	Fil4	01MAY1996	85
8	1000015	12MAY1959	44	Male	Single		0	Fil2	01FEB1999	52
9	1000016	11FEB1967	36	Male	Married		0	Fil2	01FEB2001	28

	Customer ID	Customer has any accounts	Number of Accounts	All Accounts Balance Sum	Average Interest	Loan Balance Sum	Saving Account Balance Sum	Funds Balance Sum	Loan Balance Proportion	Saving Account Balance Proportion	Funds Balance Proportion	Customer has any leasing contract	Number of leasing contracts	Totals leasing value	Total annual leasingrate
1	1000002	1	2	3100.84	5.0	1550.42	1550.42	0.00	50.00	50.00	0.00	1	1	521763.0	254.69
2	1000005	1	1	3775.31	6.0	0.00	3775.31	0.00	0.00	100.00	0.00	1	1	855215.0	232.52
3	1000006	1	1	2376.43	2.0	0.00	0.00	2376.43	0.00	0.00	100.00	1	1	560362.0	167.37
4	1000007	1	2	3625.44	5.0	0.00	1812.72	1812.72	0.00	50.00	50.00	1	2	1735708	168.75
5	1000008	1	1	3350.65	2.0	0.00	0.00	3350.65	0.00	0.00	100.00	1	1	5276.00	109.15
6	1000009	1	3	3575.46	4.0	1191.82	0.00	1191.82	33.33	0.00	33.33	1	2	591963.0	170.14
7	1000014	1	2	3000.92	4.5	0.00	3000.92	0.00	0.00	100.00	0.00	1	1	564728.0	92.51
8	1000015	1	1	2801.09	5.0	0.00	2801.09	0.00	0.00	100.00	0.00	1	1	393984.0	189.54
9	1000016	1	2	3325.66	1.0	0.00	1662.83	0.00	0.00	50.00	0.00	0	0	0.00	0.00

	Customer ID	Customer has any call center contact	Number of call center contacts	Number of complaints	Percentage of complaints	Currenty Value Segment	Last Value Segment	Change in Value Segment	Customer cancelled
1	1000002	0	0	0	.	2. SILBER	2. SILBER	1.00	0
2	1000005	0	0	0	.	3. BRONCE	3. BRONCE	1.00	0
3	1000006	0	0	0	.	3. BRONCE	3. BRONCE	1.00	0
4	1000007	0	0	0	.	2. SILBER	1. GOLD	0.00	0
5	1000008	1	1	0	0.00	2. SILBER	2. SILBER	1.00	0
6	1000009	0	0	0	.	3. BRONCE	4. LEAD	0.00	0
7	1000014	1	2	2	100.00	3. BRONCE	1. GOLD	0.00	0
8	1000015	0	0	0	.	3. BRONCE	4. LEAD	0.00	0
9	1000016	0	0	0	.	2. SILBER	2. SILBER	1.00	0

Summary

- Data Preparation is a discipline,
  not a incommodious necessity!
- The One-Row-Per-Subject Paradigm
  - Central in data mining and predictive modeling
  - Do not stop with simple transpose, summing or averaging
  - Tricky aggregations can be the key success factor
- Predictive Modeling and Historic Data:
  Which data are you allowed to use for modeling?

Recommended Reading

Data Preparation for Analytics
by Gerhard Svolba

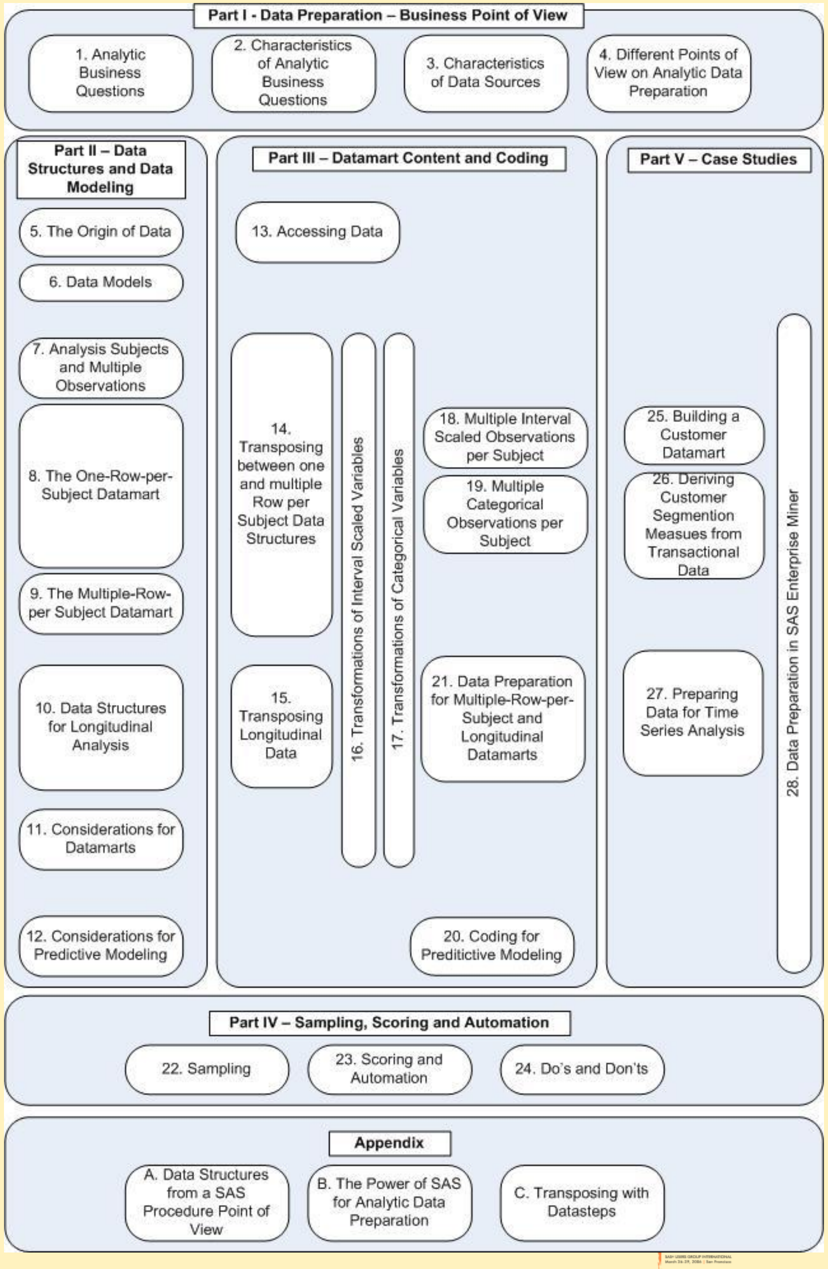
SAS-Press (#60502)

Planned publication date: October 2006

Business Rationale

Concepts

Coding Examples



Part I - Data Preparation – Business Point of View

1. Analytic Business Questions
2. Characteristics of Analytic Business Questions
3. Characteristics of Data Sources
4. Different Points of View on Analytic Data Preparation

Part II – Data Structures and Data Modeling

5. The Origin of Data
6. Data Models
7. Analysis Subjects and Multiple Observations
8. The One-Row-per-Subject Datamart
9. The Multiple-Row-per Subject Datamart
10. Data Structures for Longitudinal Analysis
11. Considerations for Datamarts
12. Considerations for Predictive Modeling

Part III – Datamart Content and Coding

13. Accessing Data
14. Transposing between one and multiple Row per Subject Data Structures
15. Transposing Longitudinal Data
16. Transformations of Interval Scaled Variables
17. Transformations of Categorical Variables
18. Multiple Interval Scaled Observations per Subject
19. Multiple Categorical Observations per Subject
20. Coding for Preditictive Modeling
21. Data Preparation for Multiple-Row-per-Subject and Longitudinal Datamarts

Part V – Case Studies

25. Building a Customer Datamart
26. Deriving Customer Segmention Measures from Transactional Data
27. Preparing Data for Time Series Analysis
28. Data Preparation in SAS Enterprise Miner

Part IV – Sampling, Scoring and Automation

22. Sampling
23. Scoring and Automation
24. Do's and Don'ts

Appendix

A. Data Structures from a SAS Procedure Point of View

B. The Power of SAS for Analytic Data Preparation

C. Transposing with Datasteps

Questions and Contact

- Gerhard Svolba (PhD)
- Email: gerhard.svolba@aut.sas.com
- Post address:
  - SAS-Austria
  - Mariahilfer Straße 116
  - 1070 Wien
  - Austria