Data Preparation for Data Science – Ingredients for a successful Machine Learning Model

Gerhard Svolba

Data Scientist, SAS Austria

#datapreparation4datascience

Medium | LinkedIn | Github | SAS-Books

Youtube: DataPreparation4DataScience

Data Science Use Cases

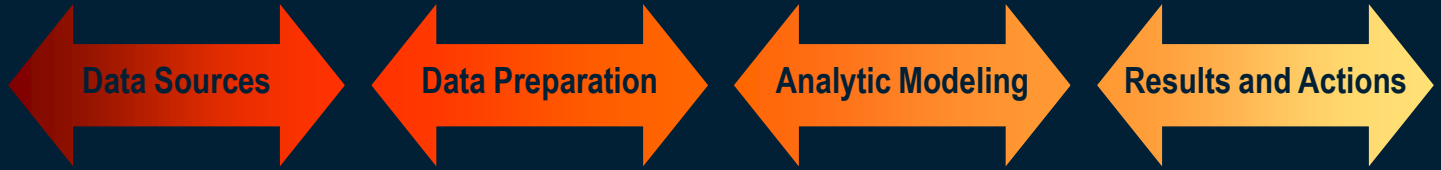
Data Preparation for Data Science

Data Assembly

Data Quality for Analytics

Feature Generation

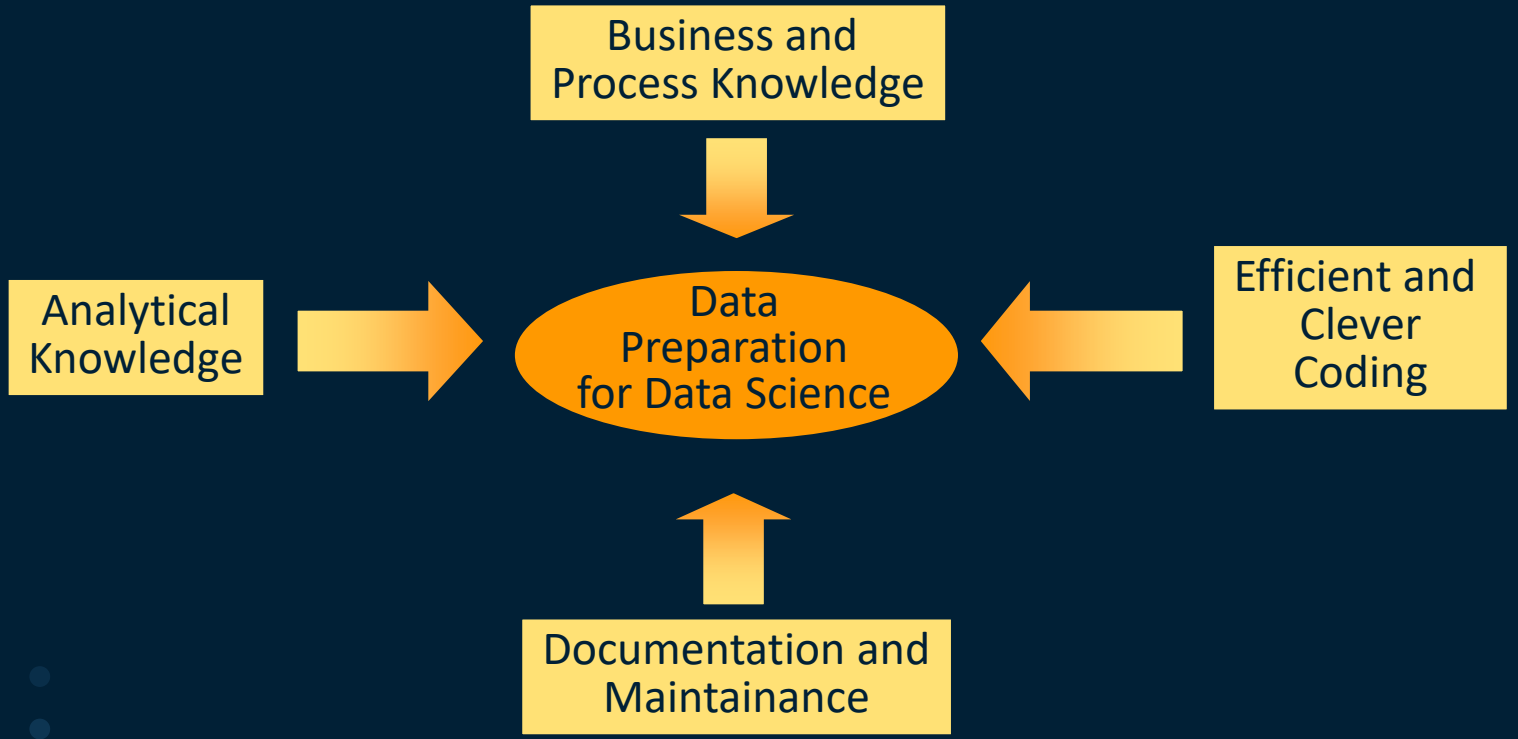
The Analysis Process: From Raw Data to Actionable Results



Data Sources	Data Preparation	Analytic Modeling	Results and Actions
Different Data Sources	Merges, Denormalization	Machine Learning Model Tuning	Profiling
Relational Models, Star Schemas Hierarchies	Transpositions, Aggregations	Predictions, Classifications, Clustering	Interpretations
	Derived Variables		Usage of Results

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

Four Dimensions for Data Preparation for Data Science



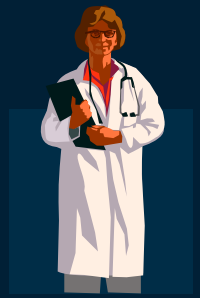
In my first AI Project, I met the following people:

I have formulated my business question. Are there any reasons, not to be back with the results in 2 days?

Simon
Retention Manager

Data Preparation for Data Science

I would like to have an analytic database with a high number of attributes and an environment to perform both, data management and analysis.



Daniele
Data Scientist

Name me the list of attributes and derived variables that you will use in your final model and which I have to deliver periodically.

Elias
IT Expert

Who is right?

- Performing Statistical Data Management
- Running Analysis within Half a Day or Four Weeks
- Statistical and Non-statistical Explanation
- Not All Business Questions Can Be Answered within a Day
- “Old Data” and Many Attributes
- When Is an Attribute Really Important?
- Automation and Clear Interfaces in Data Acquisition Make Sense

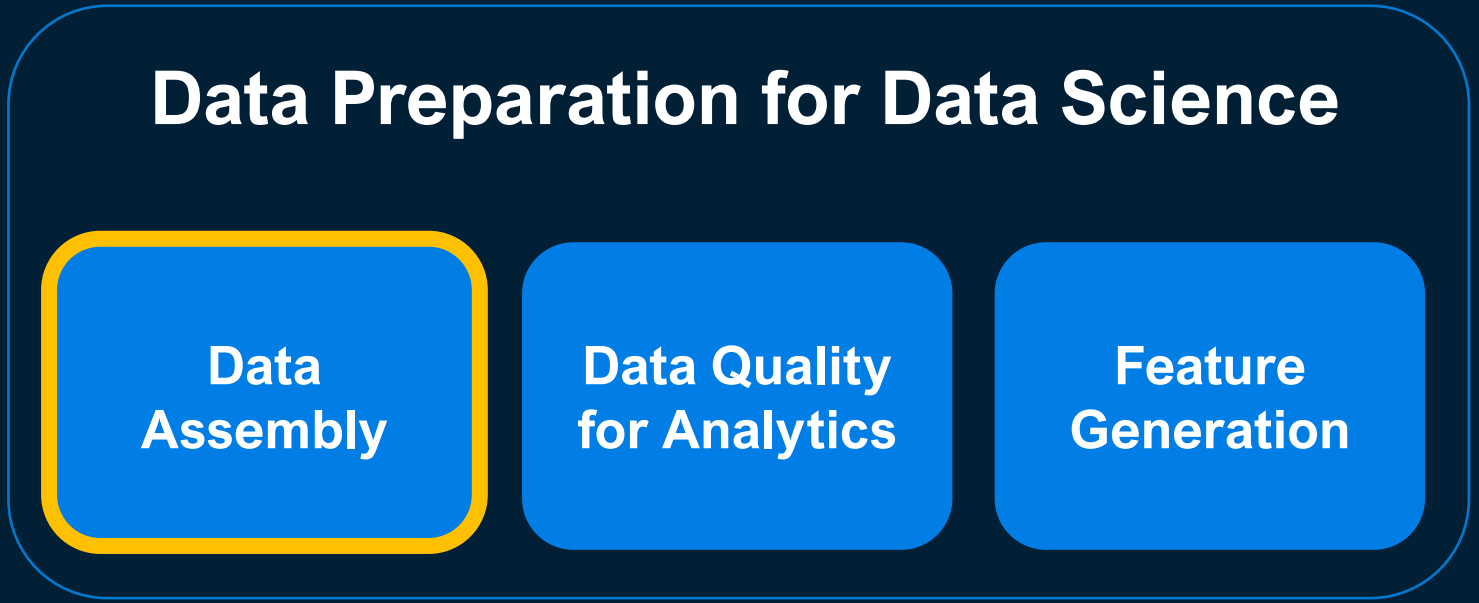
Can you build a machine learning model that predicts the cancellation risk of our customers?

- What do you mean by “cancel”?
  - Do you mean the full cancellation of the product or a decline in usage?
  - Do you want to include customers that have canceled the product but have started to use a more (or less) advanced product?
  - Do you also want to consider customers who did not themselves cancel but were canceled by our company?
- What is the business process for contacting customers?
  - Which latency period should we consider between the availability of the scores and the execution of the marketing campaign?
  - What additional attributes and explanations to you need?

Availability and alignment in the time dimension

- Historic data
  - are data that are not only available for the current period but also for historic periods, like the number of transactions for a credit card account for the last 12 months. Examples for such an intervention include marketing campaigns or medical treatments.
- Historic snapshots of data extend the concept of historic data.
  - Not only are the data from previous periods available but the snapshot of the data at a point in time in the past is also available.
- Periodic (future) availability of data
- Alignment of data.
  - The requirement of historic data can also be extended to the need to have data available that can be aligned to each other.
    - The series must cover the time period in question.
    - Historic data that are collected from different domains must have time stamps that allow aligning the available information together.

Data Preparation for Data Science



Main Types of Analytic Data Sturctures

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

Longitudinal Data Mart

	CUSTOMER	TIME	PRODUCT
1	0	0	hering
2	0	1	corned_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	corned_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

	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

The One-Row-Per-Subject Paradigm

Analysis Subject Master Table

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

Copy Variables
Create Derived Variables

Multiple Observation per Analysis Subject

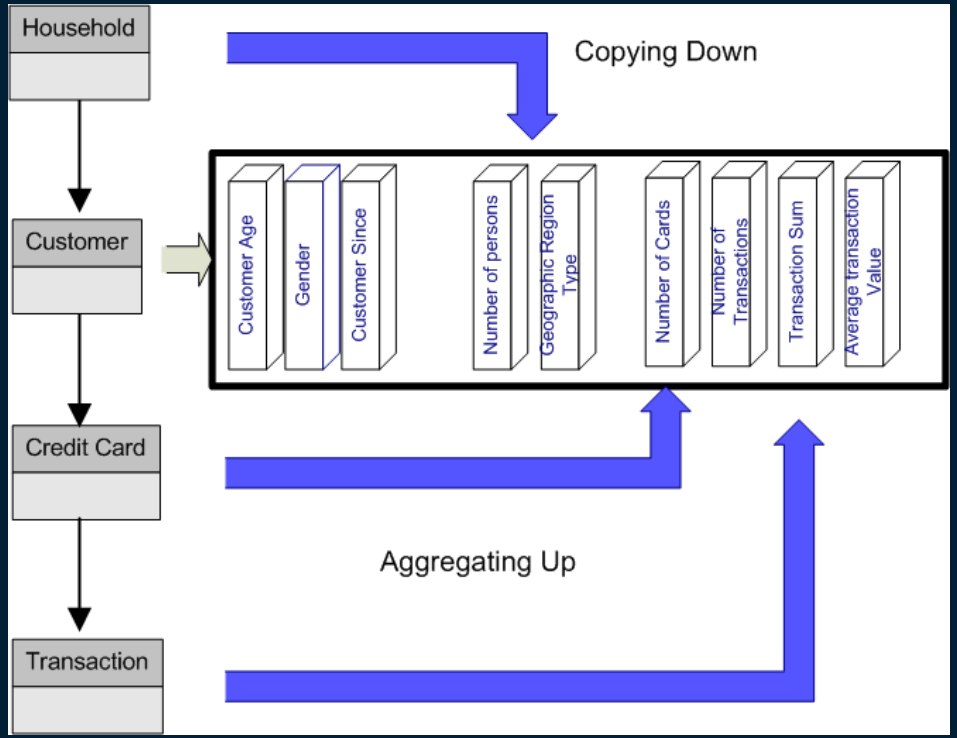
ID	Month	Type	Billing	Usage	...
1					
1					
1					
2					
2					
3					
3					
3					
4					
4					
4					
4					

Aggregate, Transpose Data
Describe Behaviour

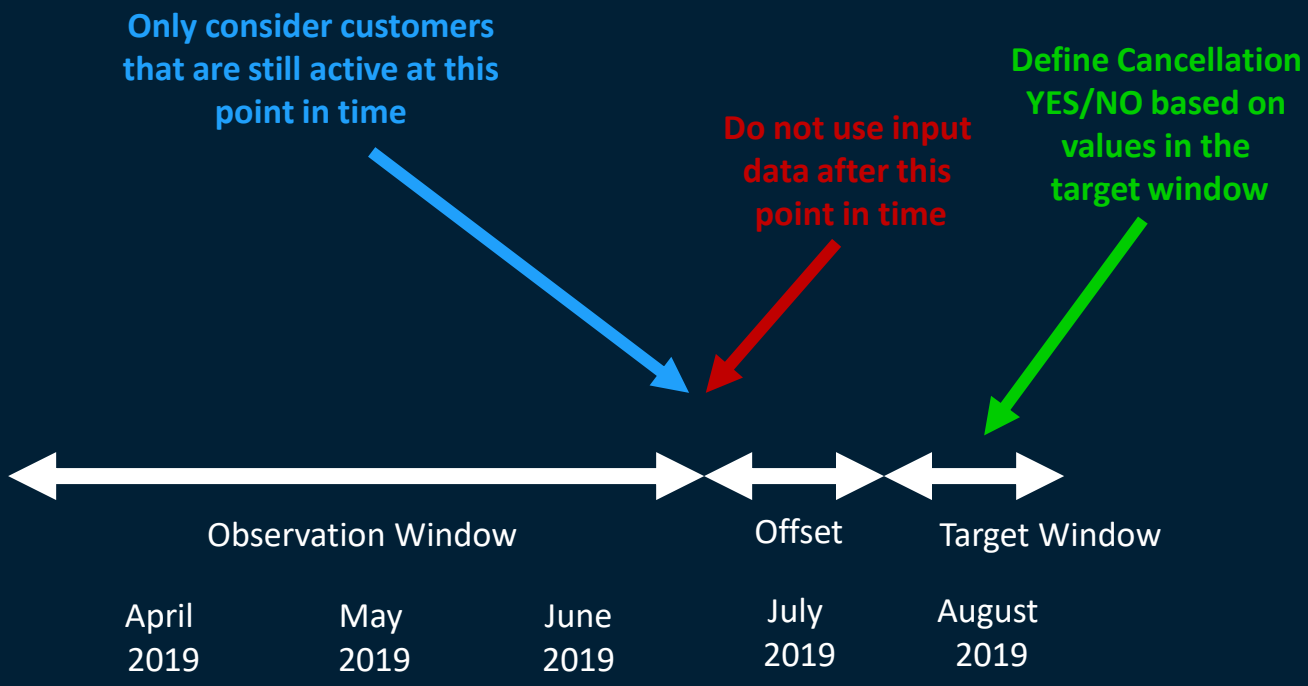
Analsis Data Mart

ID	Birth	Sex	Region	Age	...	Billing_Sum	Billing_Mean	Usage_Sum	Usage_Trend	Usage_Variab	N_Trx
1											
2											
3											
4											

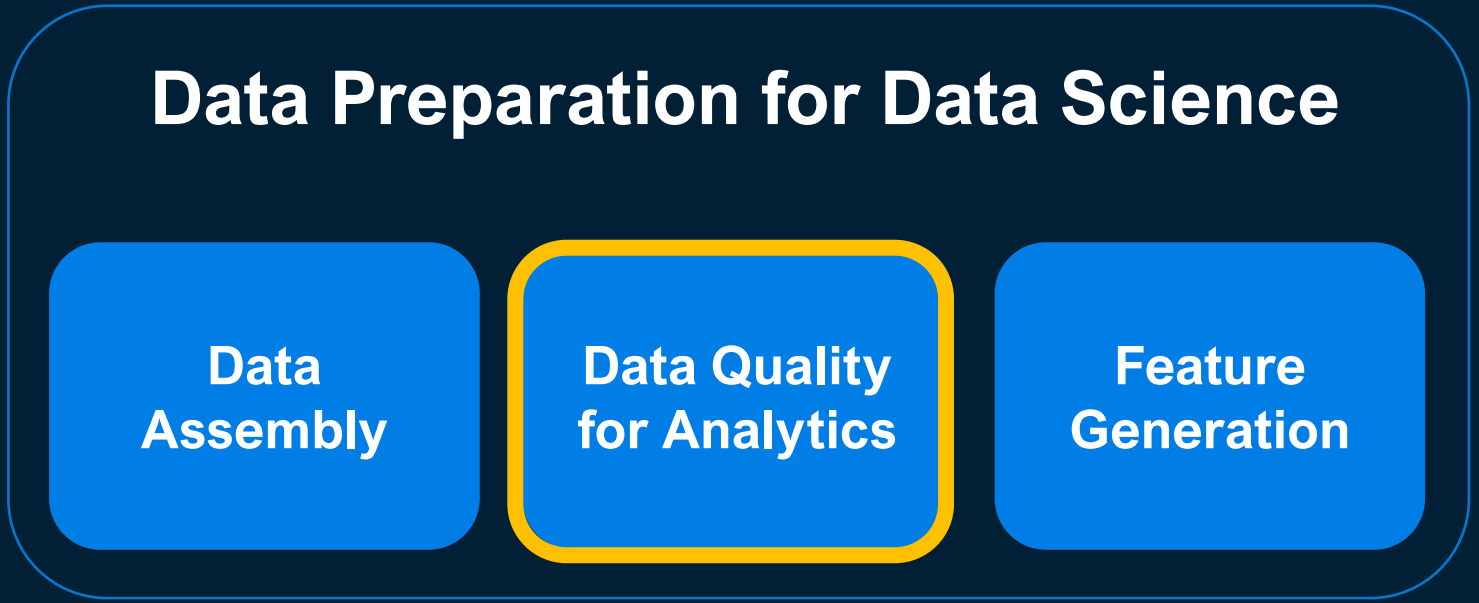
Hierarchies:
Aggregating Up + Copying Down



Considerations for Supervised Machine Learning Models: Alignment on the Time Axis



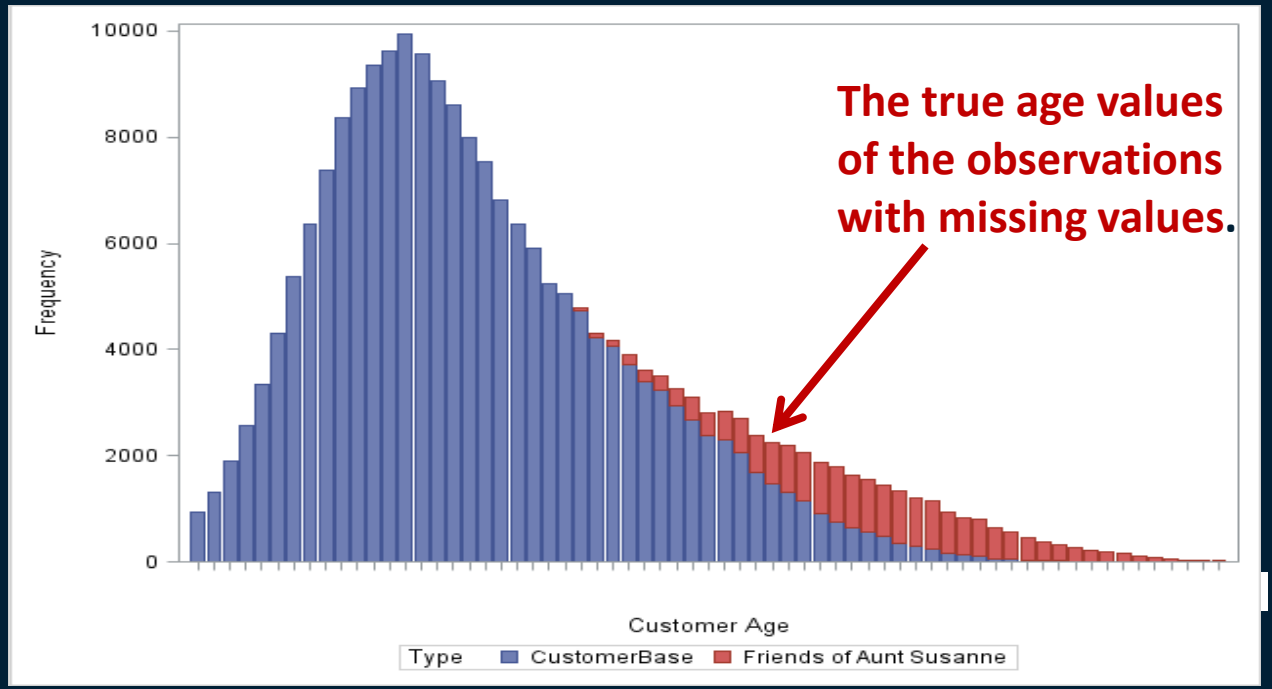
Data Preparation for Data Science



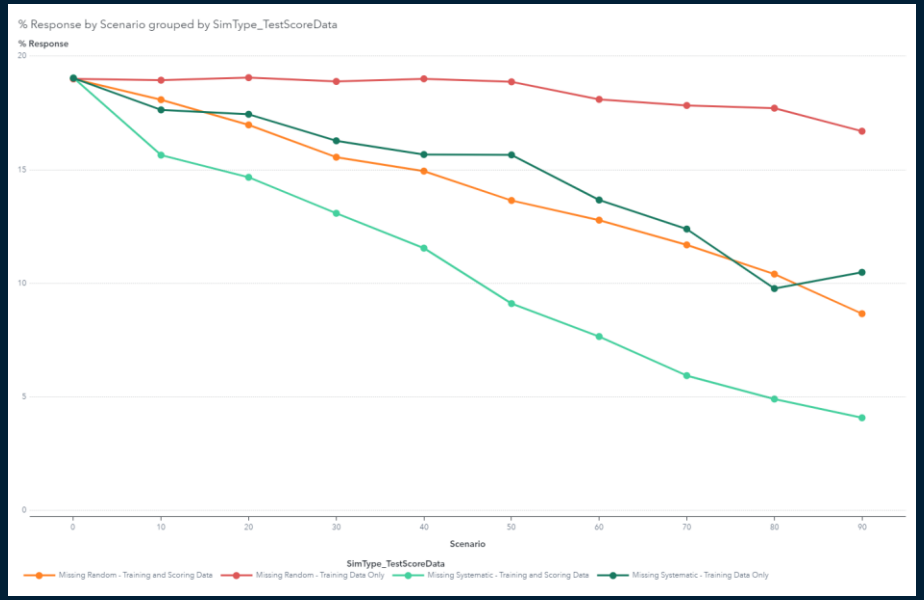
Why my Aunt Susanne gives data scientists a hard time …

- She got her phone in the mid 1960s.
- Customers‘ „Date of birth“ was of no interest at that time.
- Since the mid 1990s it is mandatory to provide the date of birth on a new contract.
- She never changed her contract type or answered any customer questionnaires.
- She is not the only one with this „data history“.

What does her phone provider see, when he looks at the customer age variable



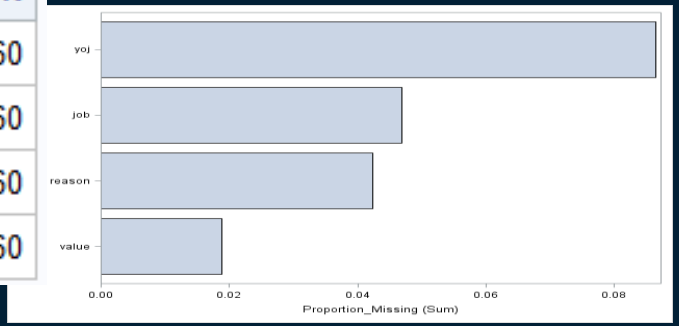
Results from simulation studies for the effect of bad data quality on model accracy



- Random missing values in training data only have limited effect.
- Missing values that occur also in the scoring data have a larger effect
- Systematic missing values have a much larger effect in general
- Takeaway:
  Not only discuss the “acceptable percentage of missing values” in your data.
  Discuss the WHY they are missing and whether this also occurs in scoring.

Typically, missing values are analyzed in a univariate way

Variable	Frequency_Missing	Proportion_Missing	N
YOJ	515	8.64%	5960
JOB	279	4.68%	5960
REASON	252	4.23%	5960
VALUE	112	1.88%	5960



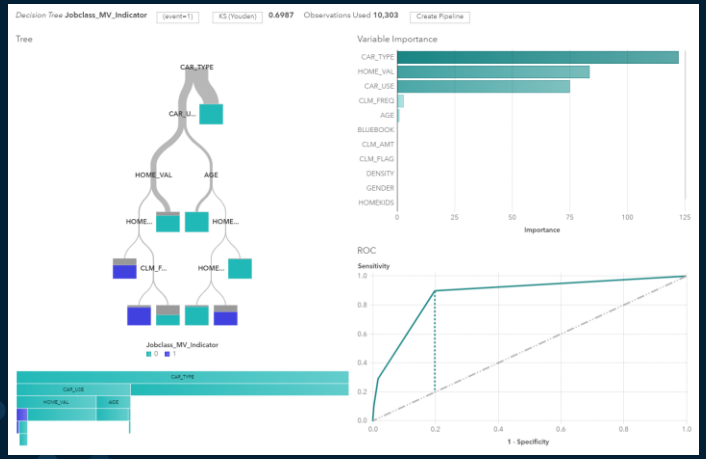
- How many of your variables are infected by the "missing value disease"?
- Not: How many "Full-Records" do you have?
- Not: Is there a pattern in the structure of missing data?

How can you detect such situations?

- Simple frequencies per variable do not help.
- Create an indicator variable „Missing YES/NO“ and compare the distribution of other variables like customer start date, product portfolio, …
- Business and process knowledge about the company is key!
- Define imputation rules based on expert knowledge.

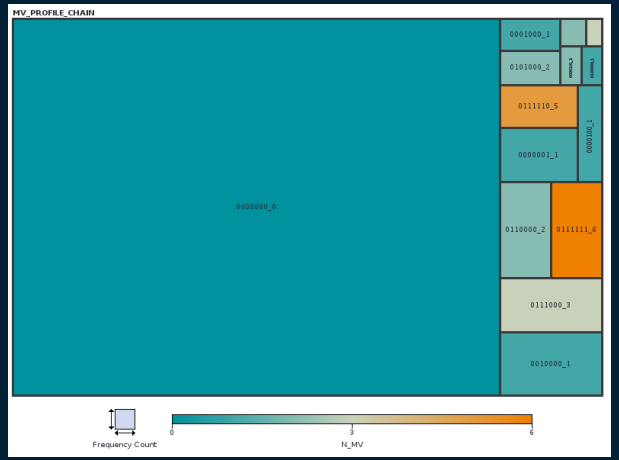
Get a deeper look into the structure of your missing values

SAS®Visual Analytics provides insight about the nature of your missing values



(Available on Youtube early April)

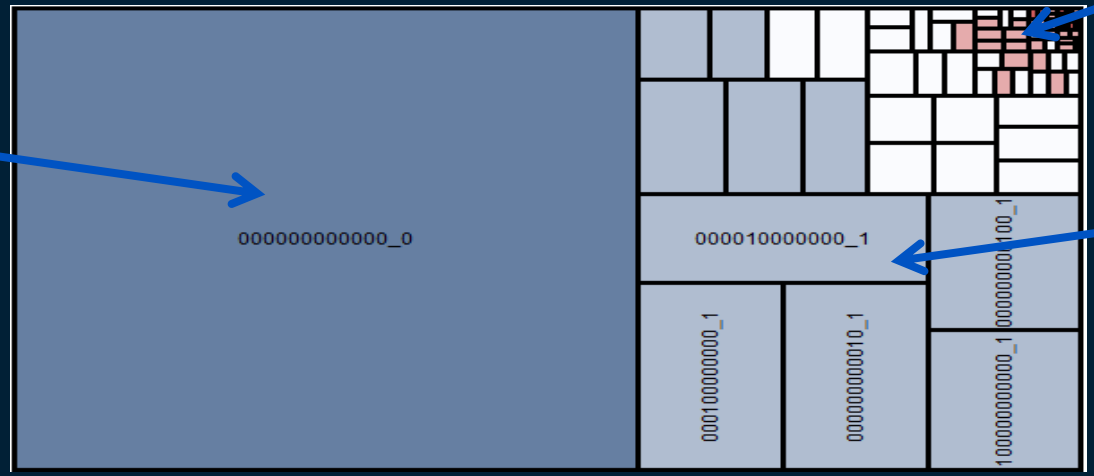
SAS Macro %MV_PROFILING to detect pattern in your missing values



The structure of MISSING VALUES in your data – get a clearer picture with the %MV_PROFILING macro

Profiling the pattern of missing values with the %MV_PROFILING macro

- Concatenate each “Missing-Value” Indicator to a string. E.g: 00100100



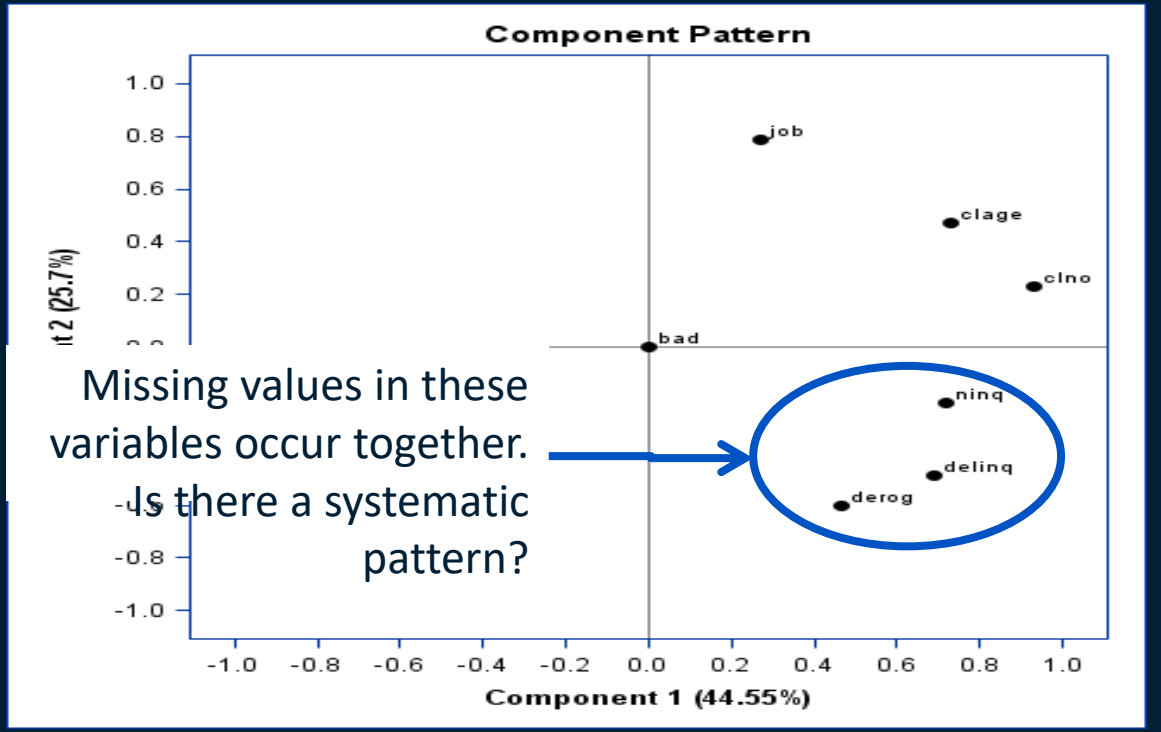
Full Records

Records with > 4 missing values

Records with a missing value in one variable

- Macros can be downloaded from #github

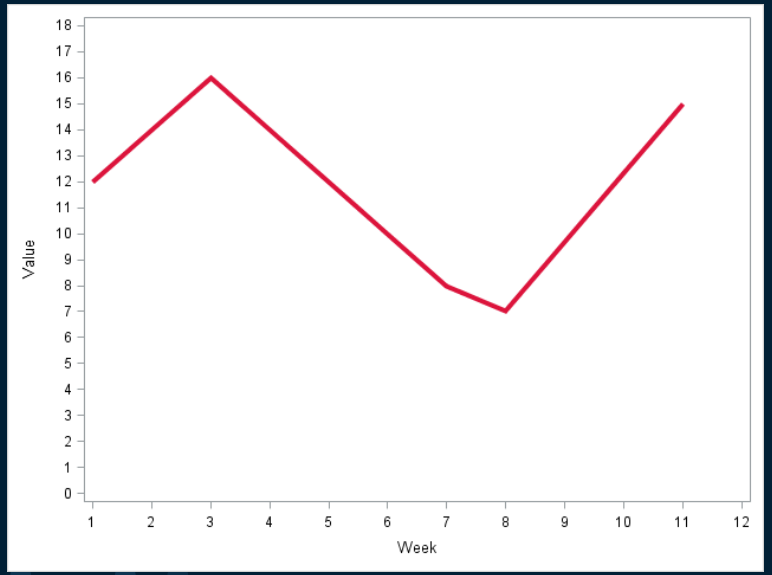
Multivariate analysis uncovers systematic patterns

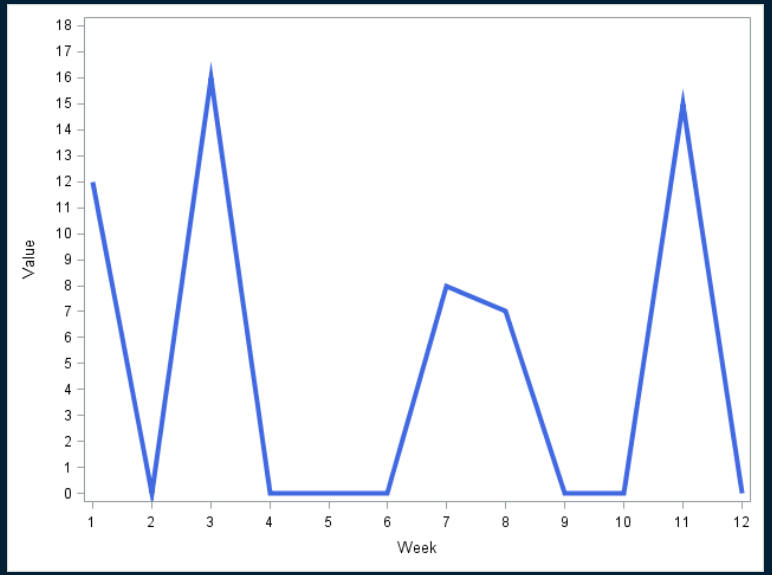


Missing values in these variables occur together. Is there a systematic pattern?

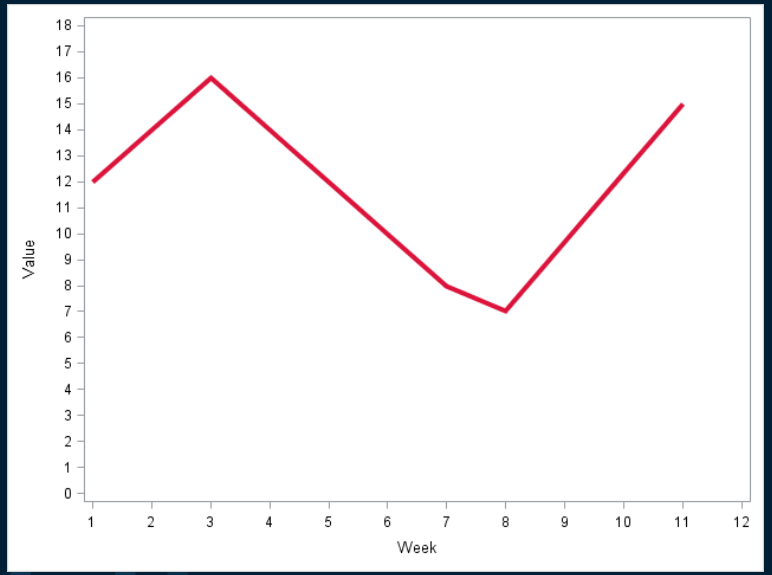
Do missing values really only matter in analytics (and not in reporting)?

Are these two graphs based on the same data?





For some measurements (inventory data) this might be the appropriate view

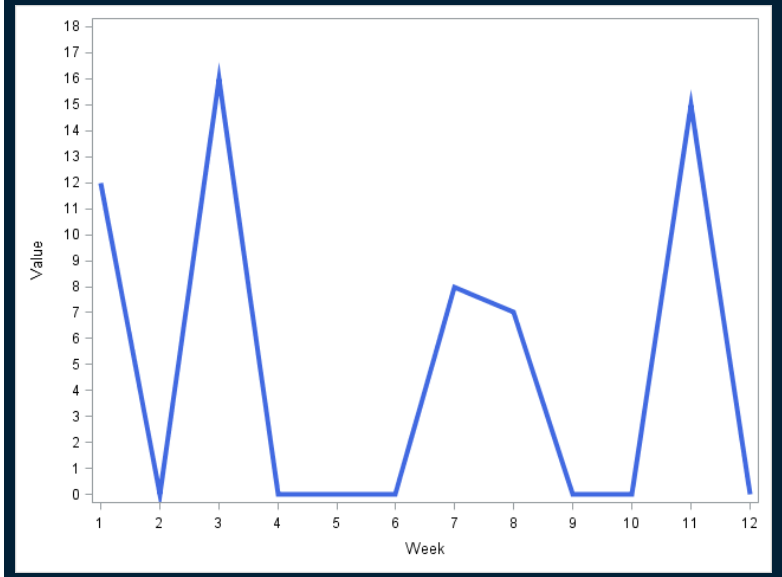


	Week	Value
1	1	12
2	3	16
3	7	8
4	8	7
5	11	15

For other measurements (movement data) this might be the appropriate view

Be careful with line-charts and missing values!

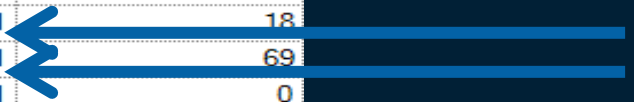
	Week	Value
1	1	12
2	2	.
3	3	16
4	4	.
5	5	.
6	6	.
7	7	8
8	8	7
9	9	.
10	10	.
11	11	15
12	12	.



Explicit or implicit missing values in longitudinal data

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

Existing Record
Value Missing



Missing Record
No Continuity

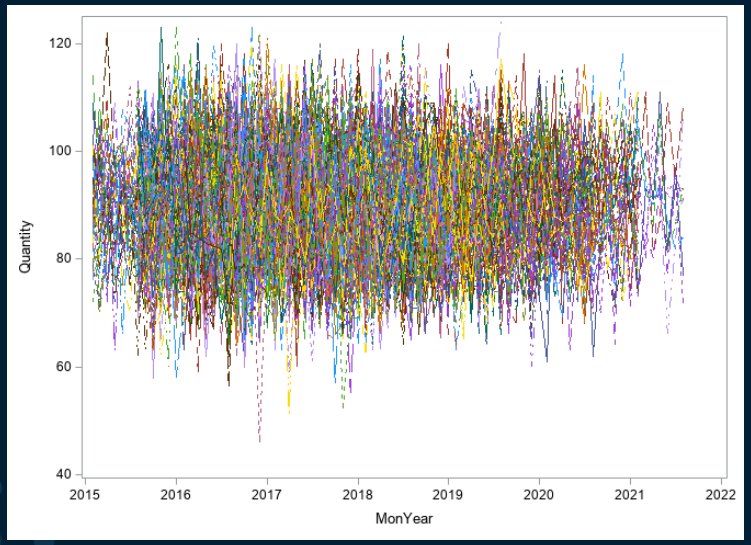
Replacing and interpolating missing values in longitudinal data with SAS

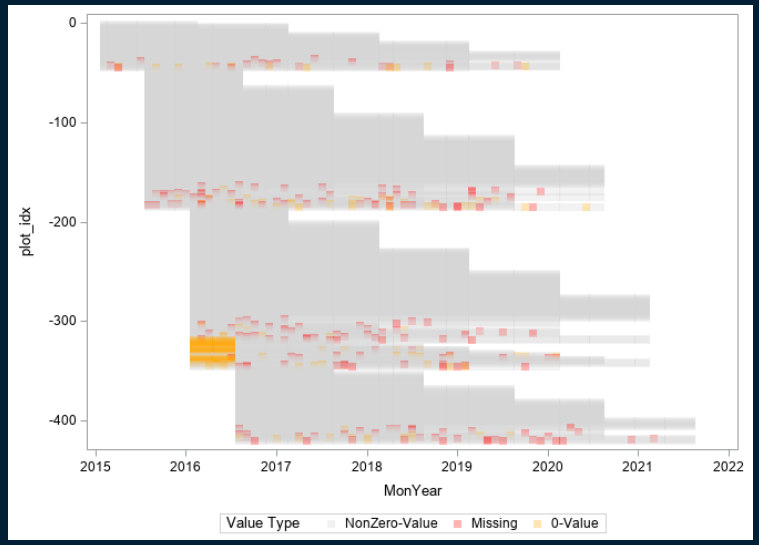
Insert missing records | Replace with 0 | Replace with last known value | Replace with mean | Interpolate based on splines

	DATE	air_mv	air_mv_zero	air_mv_previous	air_mv_mean	air_expand
1	JAN49	112	112	112	112	112
2	FEB49	118	118	118	118	118
3	MAR49	132	132	132	132	132
4	APR49	129	129	129	129	129
5	MAY49	.	0	129	284.54385965	128.29783049
6	JUN49	135	135	135	135	135
7	JUL49	.	0	135	284.54385965	144.73734152
8	AUG49	148	148	148	148	148
9	SEP49	136	136	136	136	136
10	OCT49	119	119	119	119	119
11	NOV49	.	0	119	284.54385965	116.19900978
12	DEC49	118	118	118	118	118
13	JAN50	115	115	115	115	115
14	FEB50	126	126	126	126	126
15	MAR50	141	141	141	141	141

Use PROC TIMESERIES and PROC EXPAND for these tasks!

For analytical data quality purposes: Change your View on Timeseries Data!

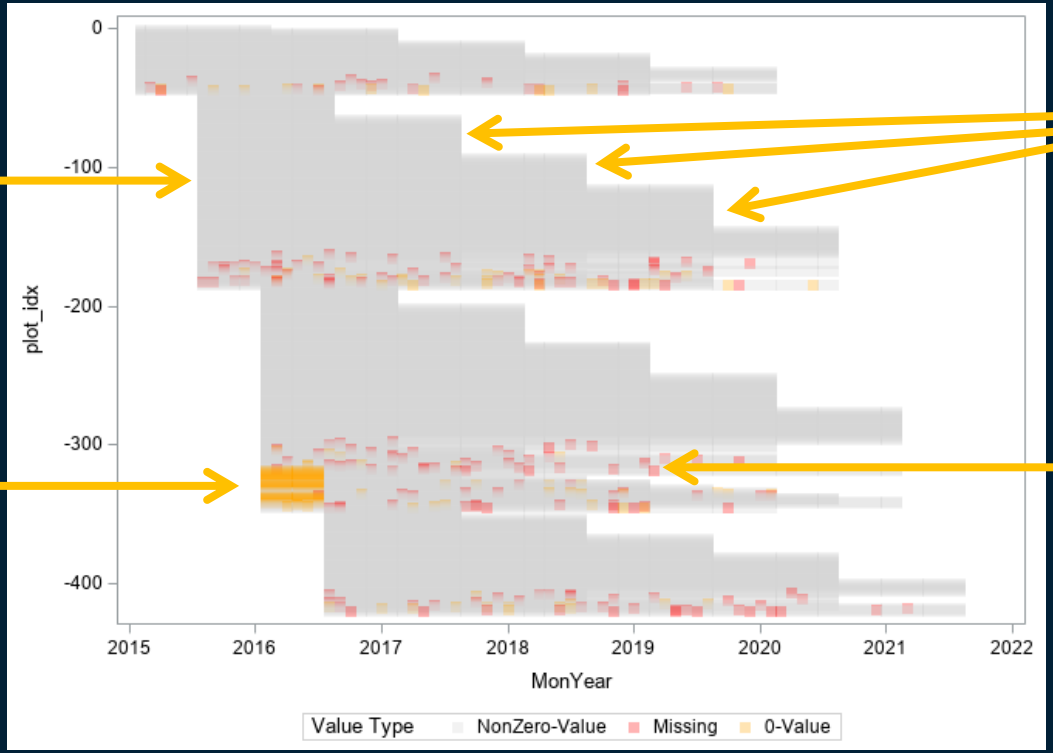




The %PROFILE_TS_MV macro provides you relevant insights

Series with full length

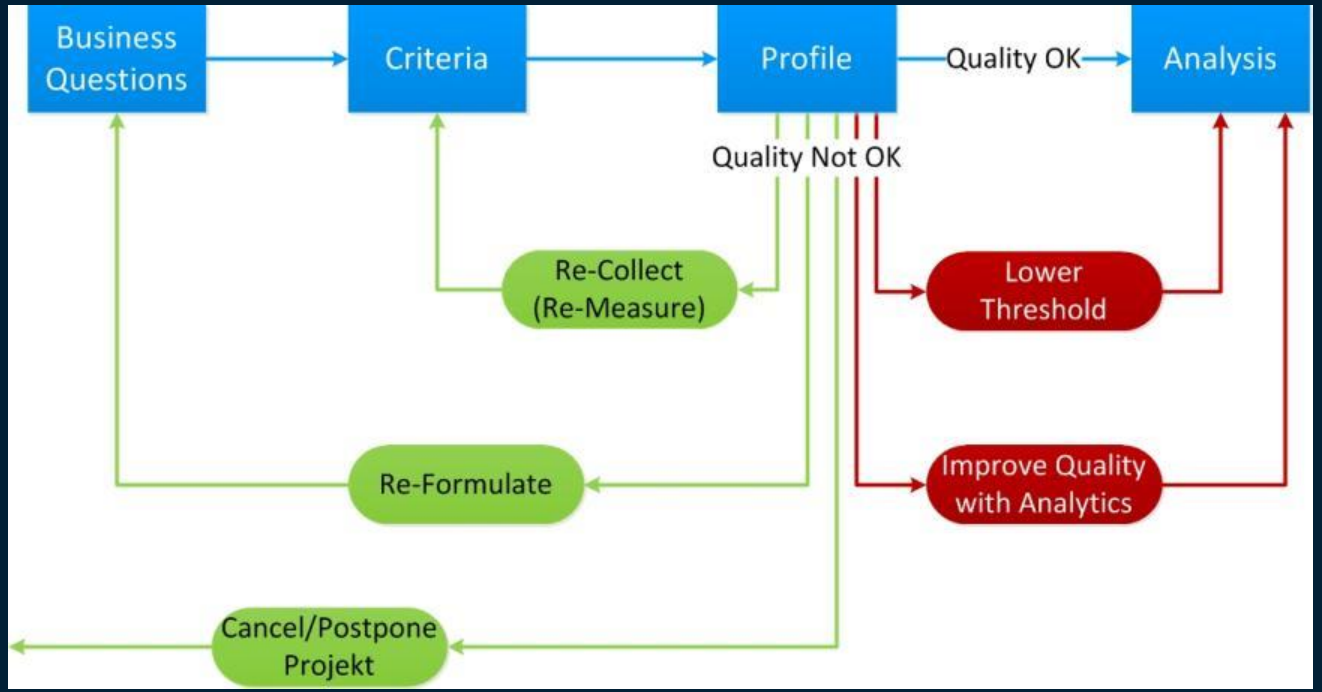
Series with leading zeros



Series with different lengths

Series with embedded missings

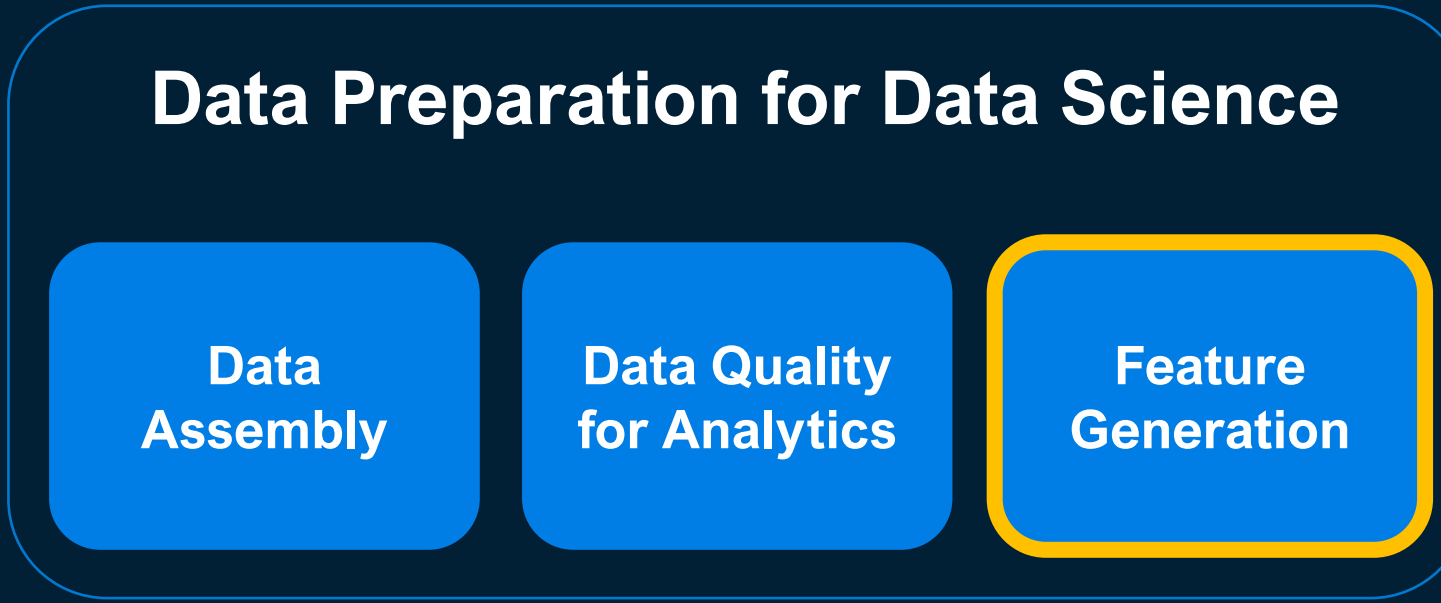
These are your options, if you learn that data quality is poor



Cost
Time, Delays
No Results

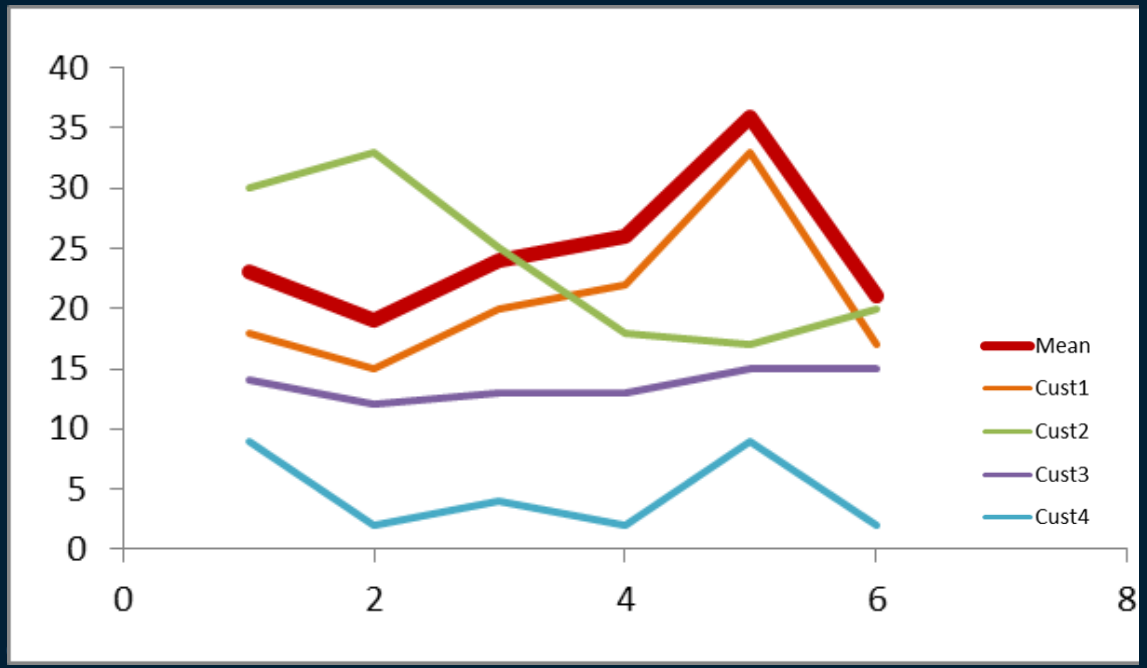
Trust
Risk of wrong decisions
Insignificance

Data Preparation for Data Science



Describe customer behaviour over time

(displaying 4 example customers and the overall mean)



Cust ID	Level
1	=-
2	=
3	-
4	--

Quick Software Demo

```
/*** Step 1
     Calculate the usage average per time interval ***/

proc sql;
 create table monthly_average
 as select month,
           mean(usage) as MonthlyAverage format = 8.2
    from usage
    group by month
    order by month;
quit;

/*** Step 2
     Merge the means per time interval to the original data ***/

proc sql;
 create table usage_enh
 as
 select a.custid,
        a.month,
        a.usage,
        b.MonthlyAverage
 from usage a,
      monthly_average  b
 where a.month = b.month;
quit;
```

```
/*** Step 3
     Calculate the the correlations between individual value and interval mean ***/

proc corr data = usage_enh
          outp=corr_usage noprint;
 var usage;
 with MonthlyAverage;
 by custid;
run;


/*** Step 4
     Rearrange to a one-row-per-subject structure ***/

proc transpose data=corr_usage out=Customer_ABT(drop=_name_);
 by custid;
 id _type_;
 var usage;
run;

data Customer_ABT;
 set Customer_ABT;
 format mean corr std 8.2;
run;
```

Feature Engineering – Be creative!

Multiple Observation per Analysis Subject					
ID	Month	Type	Billing	Usage	…
1					
1					
1					
2					
2					
3					
3					
3					
4					
4					
4					
4					

Aggregate, Transpose
Describe Behaviour

Billing_Sum	Billing_Mean	Usage_Sum	Usage_Trend	Usage_Variab	N_Trx

Interval Data

- Correlation of Values
- Course over Time
- Concentration of Values
- Seasonal Pattern

Categorical Data

- Frequency Counts
- Concatenated Frequencies
- Total and Distinct Counts

- Network Data
- Textual Data
- Images and Videos
- …

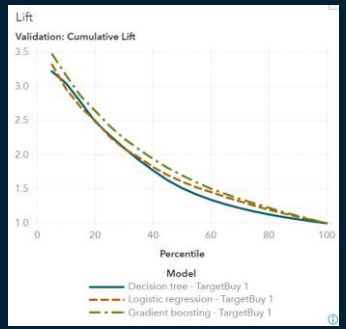
Example: Feature Engineering with Computer Vision

Aircraft Turnaround Management with SAS Event Stream Processing. Here I especially like the fact that his work allows to automatically collect data that we need as features in machine learning models and time series forecasting. How long does the unloading take place? How many minutes were spent of re-fueling?

Tracking Social Distancing on Camera Videos users computer vision to automatically calculate measure for the distance of humans in the picture. This is not just another computer vision gimmick! It allows to monitor locations over time and automatically generates measures that indicate at which places and at what time distancing rules are not observed.

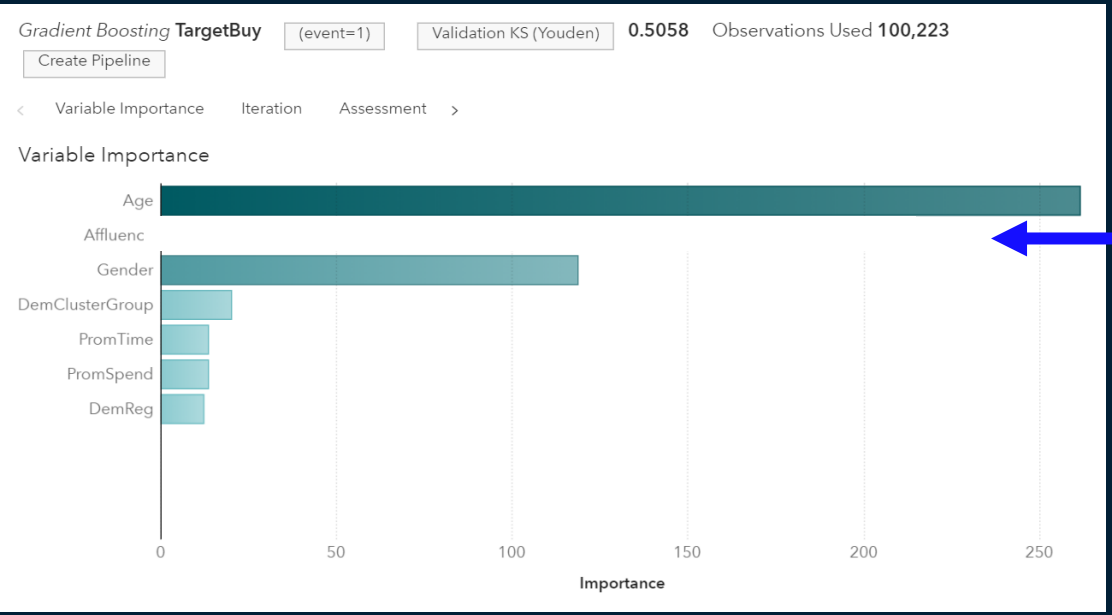
- Artikel von Michael Gorkow: https://medium.com/@michaelgorkow/track-social-distancing-using-computer-vision-2032d35cbcb4
- Aircraft Vide https://www.youtube.com/watch?v=D74rkcbDV04&t=1s
- Social Distancing Video https://www.youtube.com/watch?v=HnE9gC_ui4E
- Webinar Series von Gerhard Svolba https://medium.com/@gerhard.svolba/home-alone-e0a6bed68592

Quantify the importance of explantory variables in a predictive model with a business case



https://www.youtube.com/watch?v=Ql5oB-CDFGU&list=PLdMxv2SumIKs0A2cQLeXg1xb9OVE8e2Yq&index=5&t=0s

Variable importance chart in a gradient boosting model

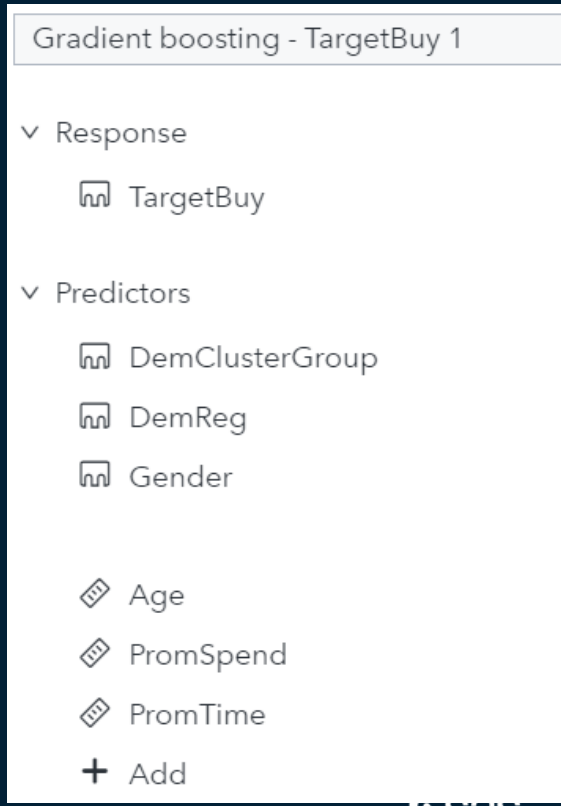


What happens, if we do not have variable „AFFLUENCE“ available?

What happens, if we do not have variable „AFFLUENCE“ available?

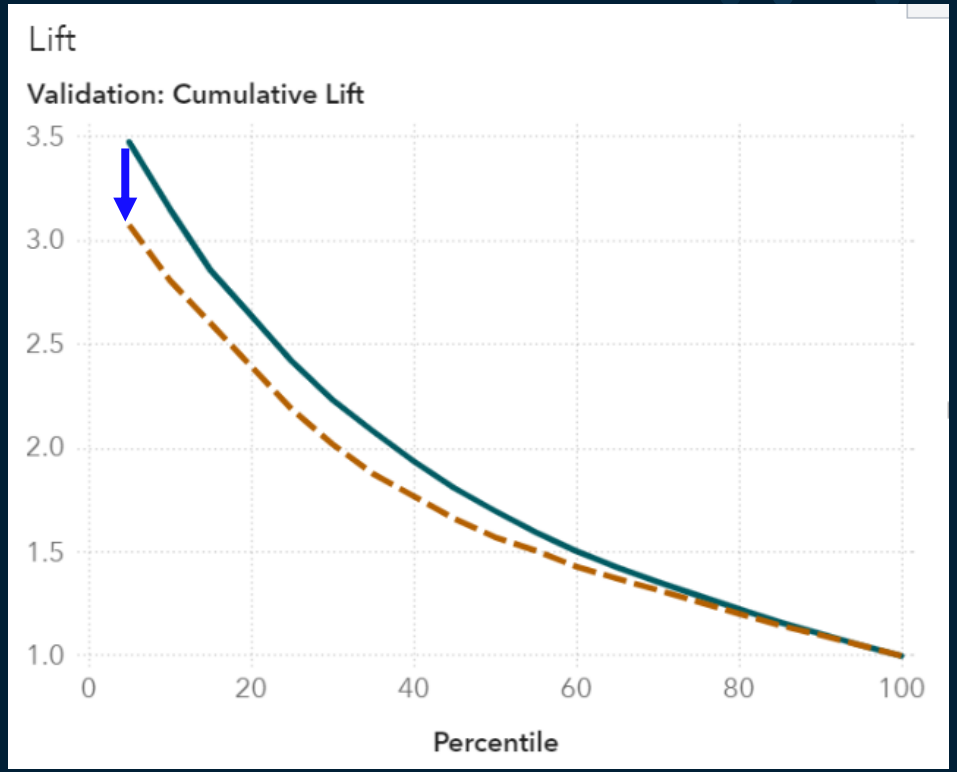
- Will other variables substitute the missing content?
- Will the model quality go down?

1. Create a copy of your model
2. Remove the variable of interest
3. Compare the old and the new model



Compare the old and the new model

- Lift drops from 3.47 to 3.07
- What does that mean in € ?

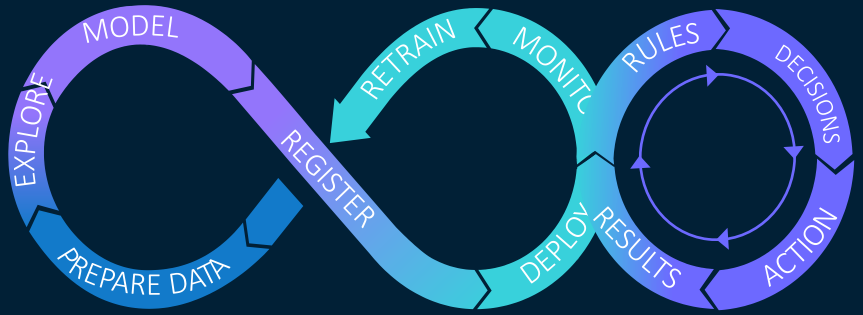


Calculating a business case

- Assume we have 2 Mio customers
- A campaign offer is sent to the top 5 % (100,000)
- A responding customer contributes a profit of € 35
- Assuming a baseline (autonomous) response of 12 %
  - A lift of 3.47 → 41.64 %
  - A lift of 3.07 → 36.84 %
- Not having variable AFFLUENCE costs us 4.8 % response
  - 100,000 * 4.8 % = 4800 missed responders * €35 = € 168,000

Conclusion

- Data Preparation is all over the analytic lifecycle!



- Data Preparation is much more than just coding!

All you need to prepare your data for data science is available in the integrated SAS Viya platform

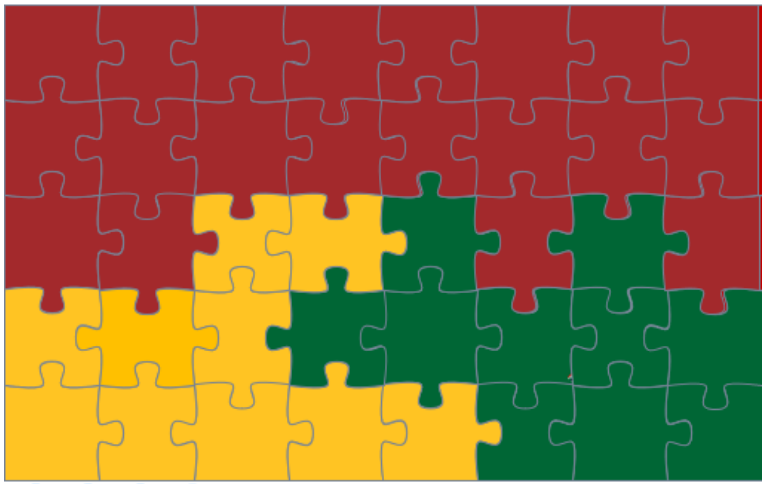
- Data Preparation / Data Quality / Feature Engineering / Variety of Analytical Methods / Visualizing Relationships / Comparing Models / What-If Scenarios / Access for different Persona Roles / Model Ops / …

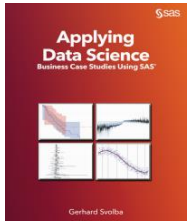
Data Preparation for Data Science

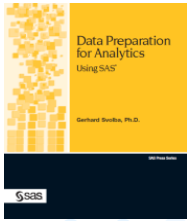
- Data Assembly
- Data Quality for Analytics
- Feature Generation

Gerhard Svolba,
Data Scientist @SAS
mailto: gerhard.svolba@sas.com

Medium | LinkedIn | Github | SAS-Books
Youtube: DataPreparation4DataScience
Data Science Use Cases







Articles and Blogs Medium LinkedIn

Webinars YouTube

Tipps & Tricks SAS COMMUNITIES

Macros & Downloads

More Links

Title	
Missing Values in Machine Learning: Why my old aunt Susanne gives data scientists a hard time	https://www.youtube.com/watch?v=5uvp4aiJxjY&list=PLdMxv2SumIKsqedLBq0t_a2_6d7jZ6Akq&index=2
“Rosetta Stone” — The most important text sample in history and the role of labeled data in machine learning	https://gerhard-svolba.medium.com/rosetta-stone-the-most-important-text-sample-in-history-and-the-role-of-labeled-data-in-a890c782344c
The structure of MISSING VALUES in your data - get a clearer picture with the %MV_PROFILING macro	https://communities.sas.com/t5/SAS-Communities-Library/The-structure-of-MISSING-VALUES-in-your-data-get-a-clearer/ta-p/712770
Using the TIMESERIES procedure to check the continuity of your timeseries data	https://communities.sas.com/t5/SAS-Communities-Library/Using-the-TIMESERIES-procedure-to-check-the-continuity-of-your/ta-p/714678
Replace MISSING VALUES in TIMESERIES DATA using PROC EXPAND and PROC TIMESERIES	https://communities.sas.com/t5/SAS-Communities-Library/Replace-MISSING-VALUES-in-TIMESERIES-DATA-using-PROC-EXPAND-and/ta-p/714806
Detecting and Treating Missing Values in Longitudinal data	https://www.youtube.com/watch?v=egG_J8lb61Y&list=PLdMxv2SumIKsqedLBq0t_a2_6d7jZ6Akq&index=1
Profiling the Missing Value Structure of Your Timeseries Data	https://www.youtube.com/watch?v=SWh_MToyn_A&list=PLdMxv2SumIKsqedLBq0t_a2_6d7jZ6Akq&index=3
Have a look at your TIMESERIES data from a bird's-eye view - Profile their missing value structure	https://communities.sas.com/t5/SAS-Communities-Library/Have-a-look-at-your-TIMESERIES-data-from-a-bird-s-eye-view/ta-p/717449