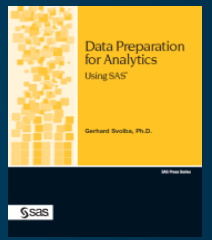
Data Science in Action – 10 Dinge, die Advanced Analytics und Data Science für Ihr Unternehmen tun kann

Gerhard Svolba, SAS Austria

Mannheim, 1. März 2018 - KSFE 2018



https://github.com/gerhard1050/

THE POWER TO KNOW.


Data Preparation for Analytics
Using SAS
Ottakringer
Stiegl
BIER
Das Salzburger Bier.
Data Quality for Analytics Using SAS
GUT BESSER
Gösser

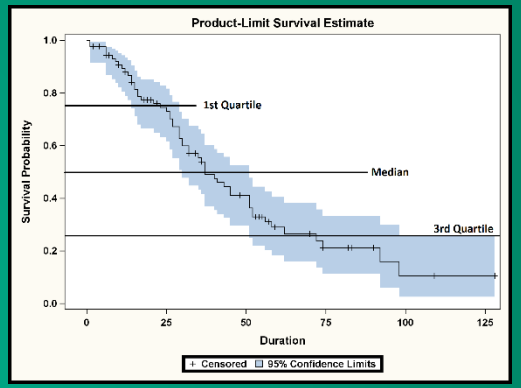

Data Preparation for Analytics Using SAS
Gerhard Svolba, Ph.D.
Applying Data Science
Business Case Studies Using SAS
Data Quality for Analytics Using SAS
Gerhard Svolba
GUT BESSER
Ottakringer
1837
Stiegl
BIER
Das Salzburger Bier.
Gösser

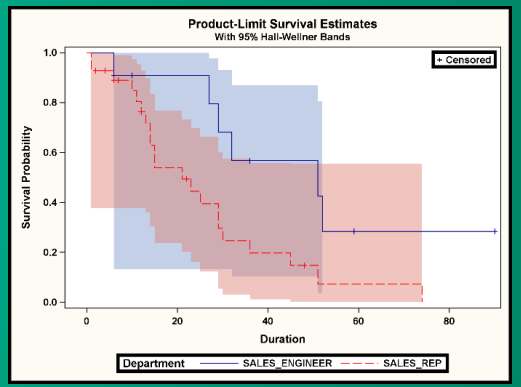
Agenda

- 10 mal „Data Science in Action“
  - Supervised Machine Learning Methoden
  - Unsupervised Machine Learning Methoden
  - Simulationen

Data Science in Action: #1

Performing Headcount Survival Analysis for Employee Retention



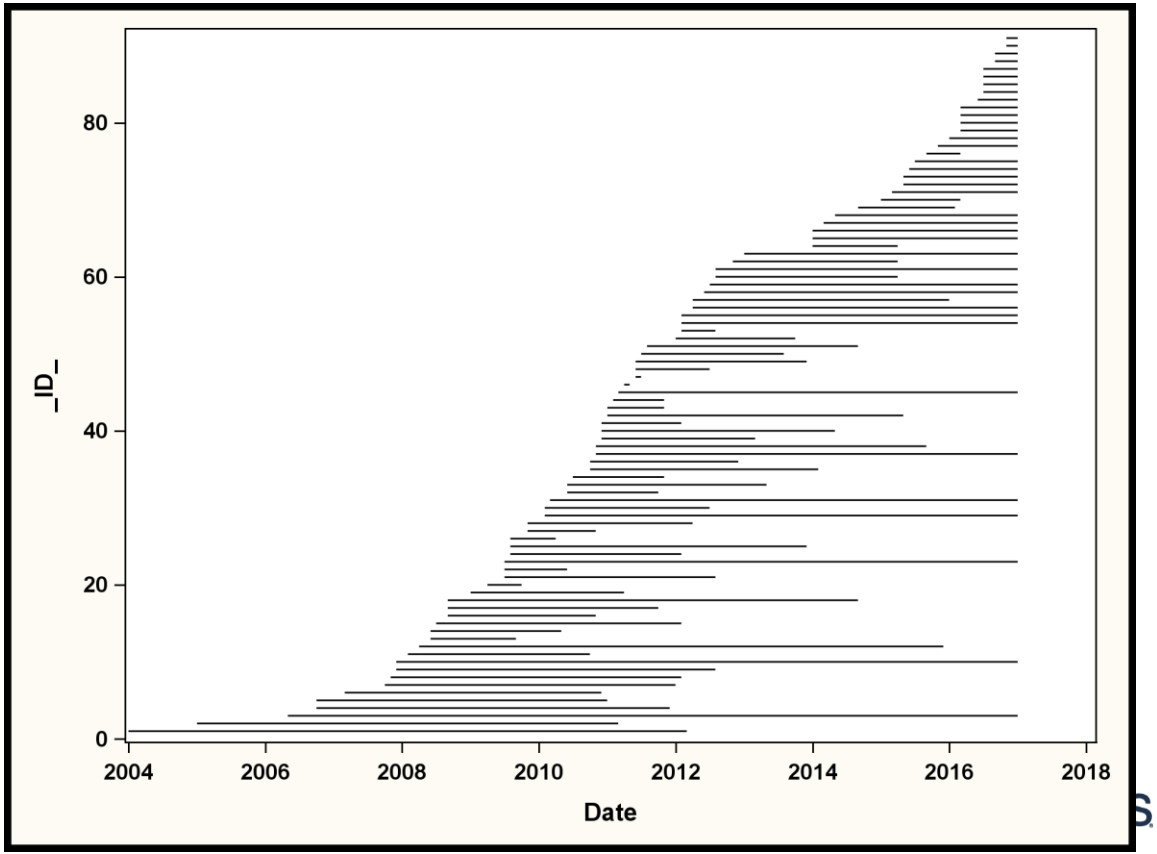


Can assumptions about the average length of time intervals be made, even if most of the endpoints have not yet been observed?

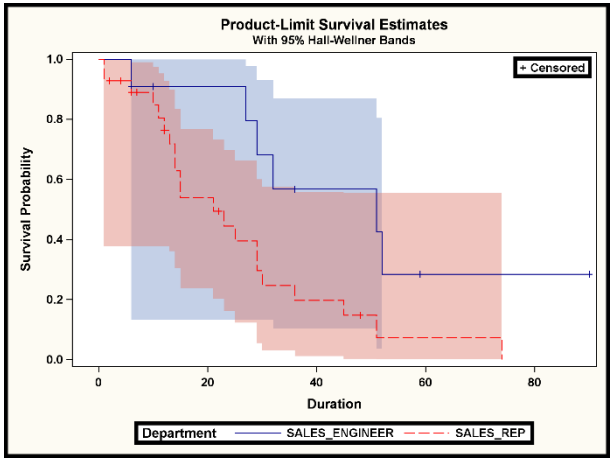
Survival analysis methods: Kaplan-Meier estimates
Cox Proportional Hazards regression
Survival Data Mining

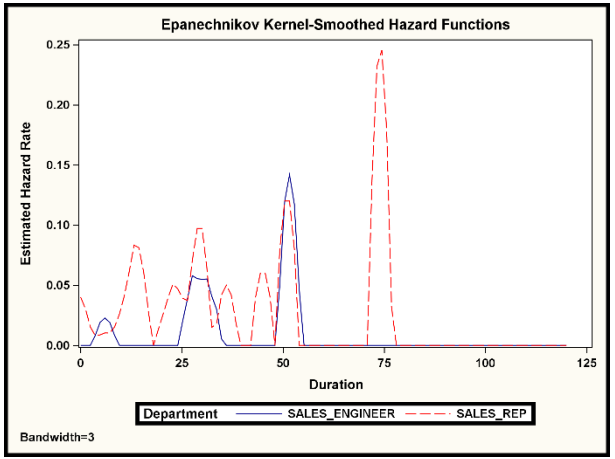
Nicht zu allen Mitarbeitern haben wir ein „Ereignis-Datum“ (Glücklicherweise)

- Betrachten der Karrieren pro Mitarbeiter
  - Unterschiedliche Länge
  - Kündigung oder „zensiert“



Die Kaplan Meier Methode und die Cox Proportional Hazards Regression verarbeitet zensierte Beobachtungen





Kaplan Meier Methods und Cox Proportional Hazards Regression:
Sales engineers haben eine bessere „survival time“ als sales representatives.

.

Betrachten der Hazard Kurven:
Es gibt ein hohes Risiko die Sales Engineers nach 26 und 50 Monaten zu verlieren.

Time-to-Event Analyse mit SAS/STAT Procedures

Proc LIFETEST und PROC PHREG

```
proc lifetest data=employees outsurv = survplot
              plots=(hazard(bandwidth=3 maxtime=120)
              survival(cb=hw));
 time duration*status(1);
 strata department;
 where department  in ("sales_rep", "sales_engineer");
run;
```

```
PROC PHREG DATA=Employees outest = ParamEstimates;
 CLASS department gender TechKnowHow StartPeriod/ PARAM=effect REF=first;
 MODEL Duration*Status(1)= department gender / SELECTION=stepwise;
 OUTPUT OUT=surv_pred survival=SurvPred
                      Atrisk  =ObsAtRsik
                      LD      =DisplacmLikelihood;
RUN;
```

„Wie lange wird Gerhard Svolba noch in unserem Unternehmen sein?“

Vorhersage der Verweildauer für indivudelle Mitarbeiter

Ausgehend von bestimmten Risikofaktoren

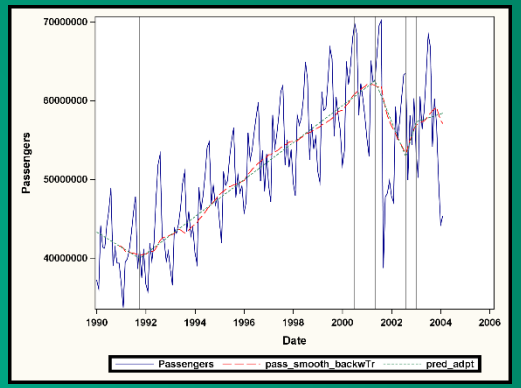
wie hoch ist die erwartete Survival in 6 Monaten

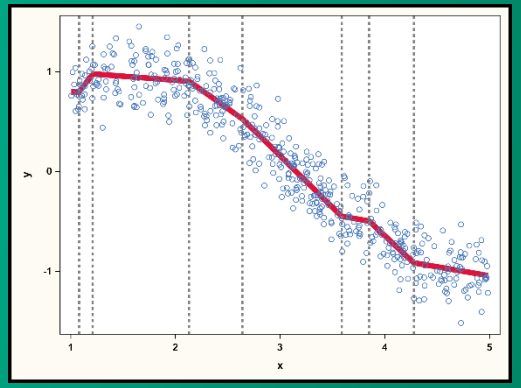
und die „Überlebens“-wahrscheinlichkeit für die nächsten 6 Monate

EmpNo	Department	Gender	TechKnowH...	_T_	EM_SURVFCST	EM_SURVEVENT	T_FCST
1003	TECH_SUPPORT	M	YES	128	0.240	0.000	134
1010	TECH_SUPPORT	M	YES	109	0.240	0.011	115
1023	SALES_ENGINEER	M	YES	90	0.108	0.313	96
1029	TECH_SUPPORT	M	YES	83	0.386	0.133	89
1031	TECH_SUPPORT	F	YES	82	0.177	0.219	88
1037	ADMINSTRATION	M	NO	74	0.471	0.066	80
1045	ADMINSTRATION	M	NO	70	0.494	0.053	76
1054	TECH_SUPPORT	F	YES	59	0.316	0.102	65
1055	SALES_ENGINEER	M	YES	59	0.313	0.103	65

Data Science in Action: #2

Detecting Structural Changes and Outliers in Longitudinal Data



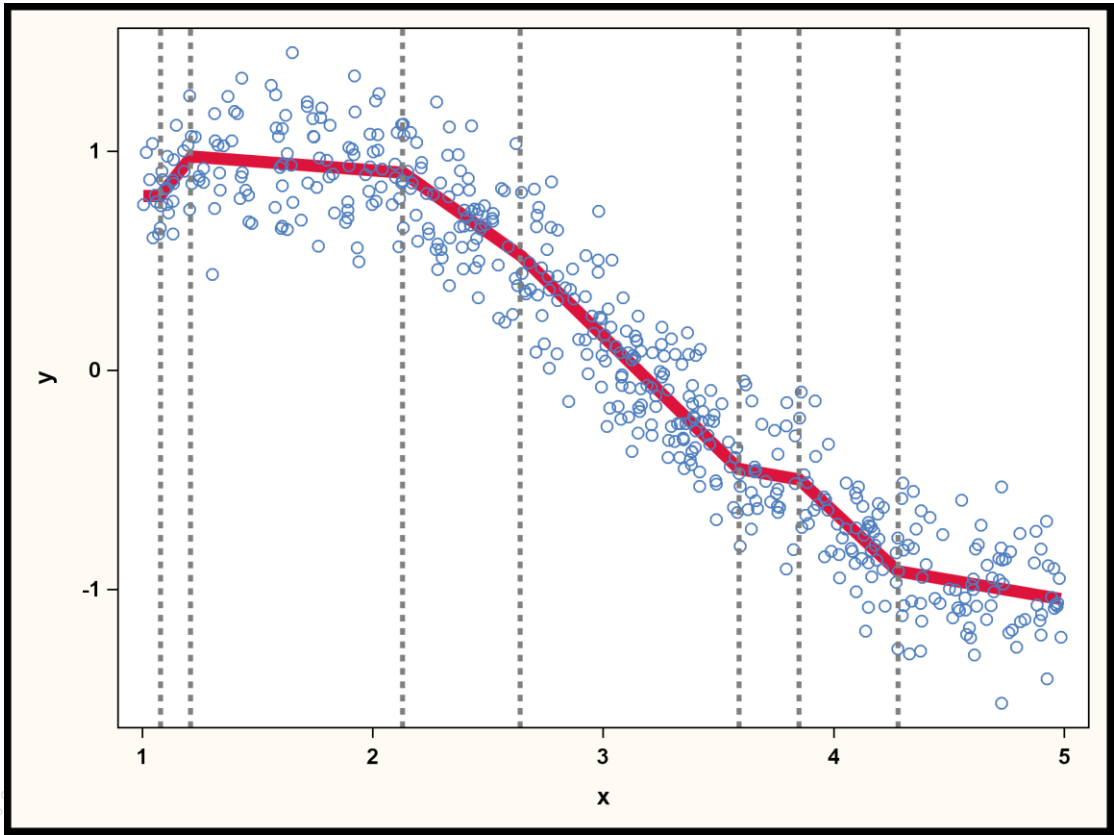


Can events and changes in the course over time be automatically detected?

Smoothing Of Longitudinal Data
Multivariate Adaptive Regression Splines
Automatic Breakpoint Detection
Automatic Detection of Outliers with ARIMA Models

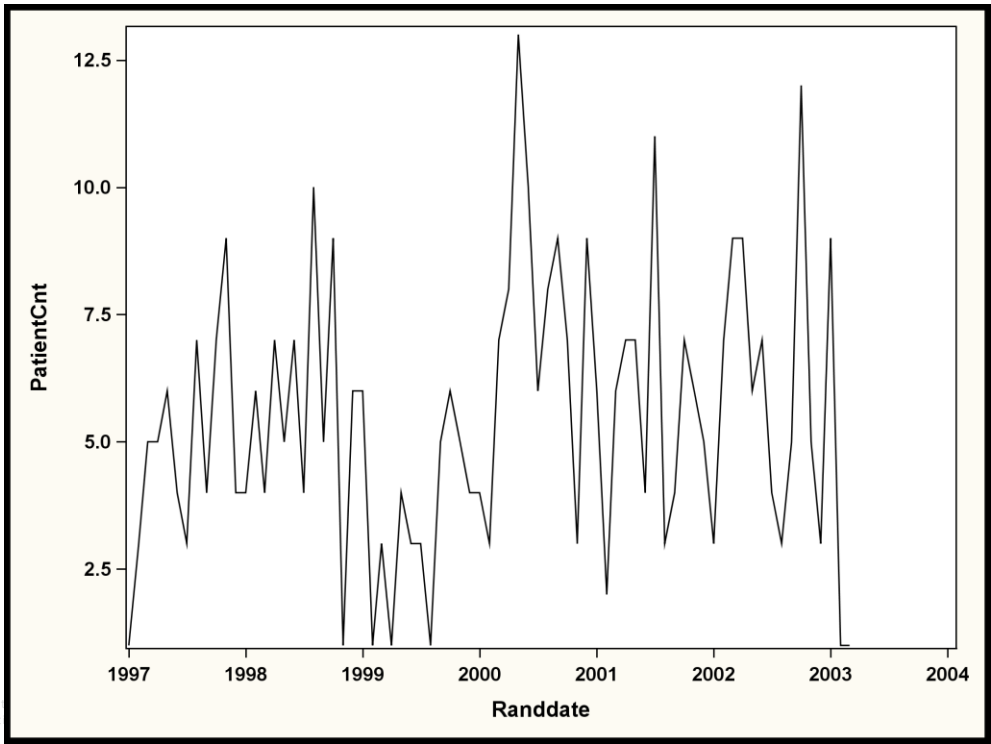
Multivariate Adaptive Regression Splines mit der ADAPTIVEREG Procedure

```
proc adaptivereg data=Demo_xy
                 plots=all
                 details=bases;
 model y = x ;
 ods output BWDParams=KnotPoints;
 output out=Demo_xy_Data_y_adpt
        predicted=pred_y;
run;
```

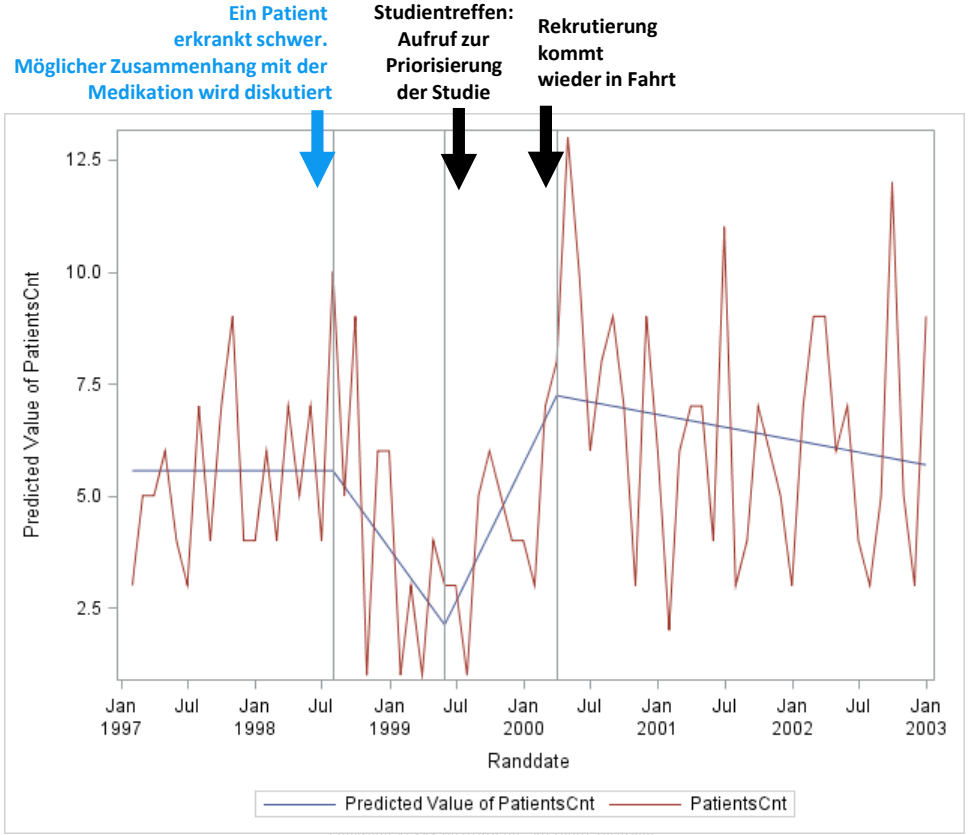


Anzahl der rekrutierten Patienten in einer klinischen Studie im Zeitverlauf

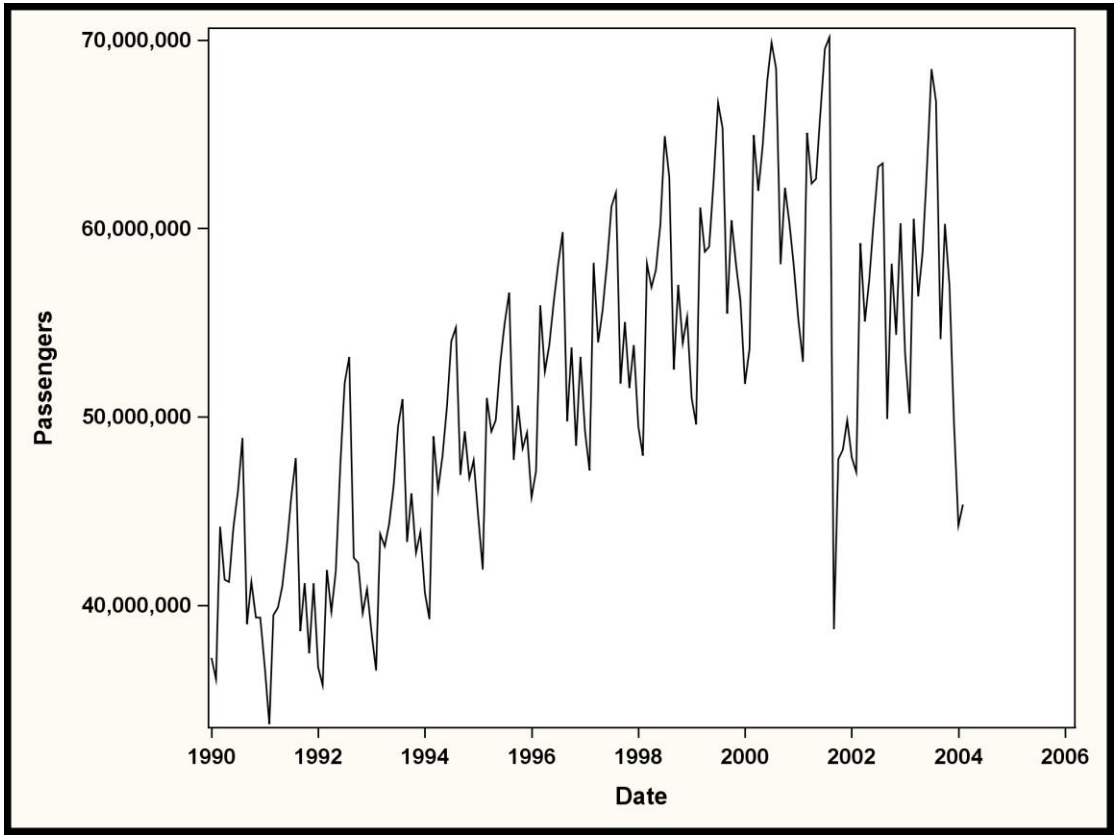
```
proc adaptivereg data=patients_recruitment plots=all;
  model PatientCnt = randdate;
  ods output BWDParams=KnotPoints;
  output out=recruit_adpt
         predicted=pred_adpt;
run;
```



Was ist zu bestimmten Zeitpunkten in meiner klinischen Studie passiert?



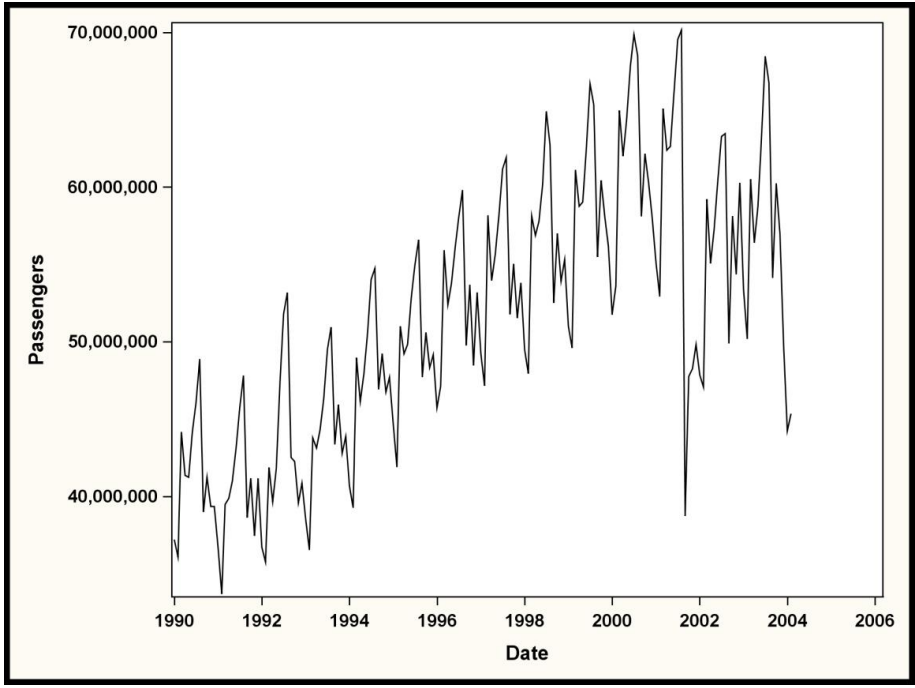
Anzahl der Flug-Passagiere in den Jahren 1990 bis 2004

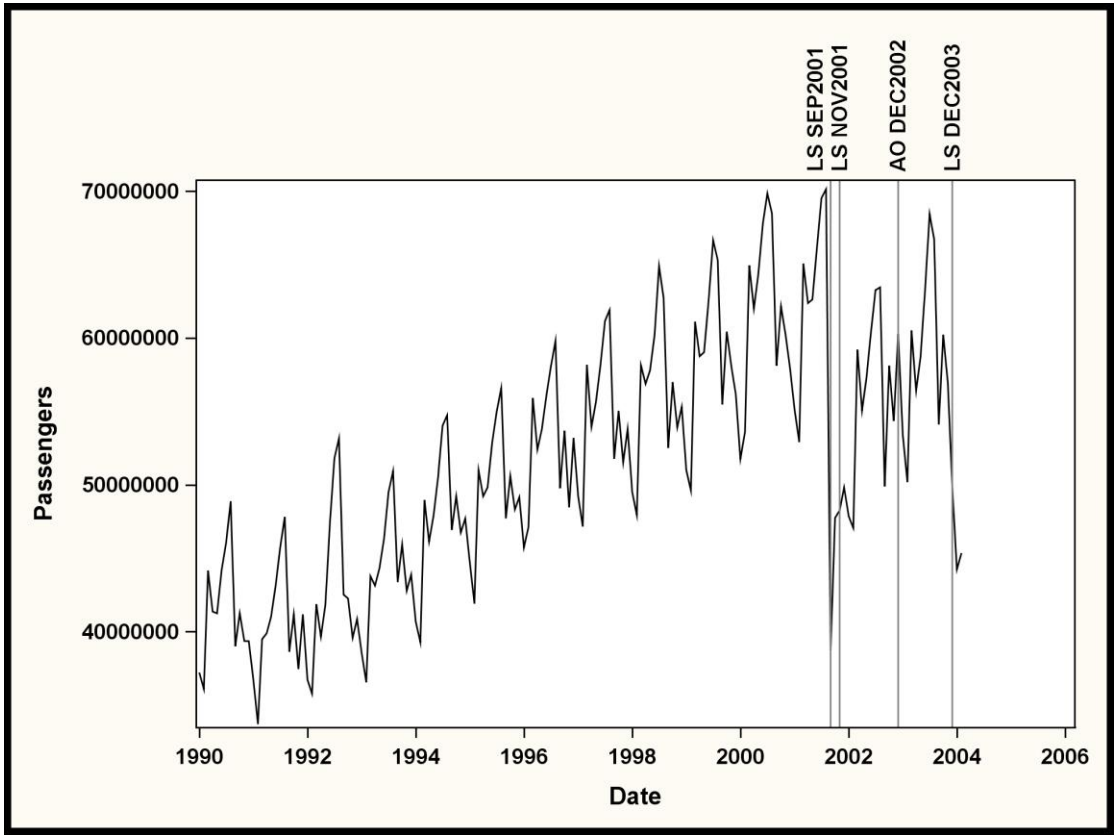


Automatische Ausreißer-Erkennung mit der X13-Procedure

```
proc x13 data=flights_911 date=date;
   var passengers;
   arima model=( (0,1,1)(0,1,1) );
   outlier;
run;
```

Regression Model Parameter Estimates For Variable Passengers						
Type	Parameter	NoEst	Estimate	Standard Error	t Value	Pr > |t|
Automatically Identified	LS SEP2001	Est	-17993818	1113414.76	-16.16	<.0001
	LS NOV2001	Est	5939640.53	1123179.20	5.29	<.0001
	AO DEC2002	Est	5039786.79	1111284.48	4.54	<.0001
	LS DEC2003	Est	-8531934.1	1249512.29	-6.83	<.0001

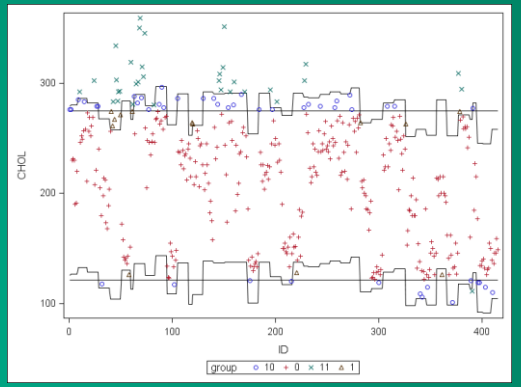


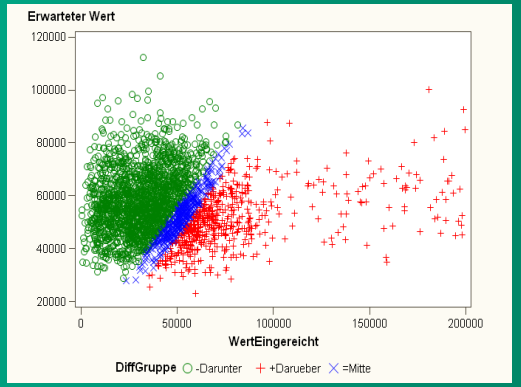

LS SEP2001
LS NOV2001
AO DEC2002
LS DEC2003
70000000
60000000
50000000
40000000
Passengers
1990
1992
1994
1996
1998
2000
2002
2004
2006
Date

Data Science in Action: #3

Proving a reference value that considers all available co-information

Can analytics help me to reduce the "Yes, but … " sentences in my business dicussions?

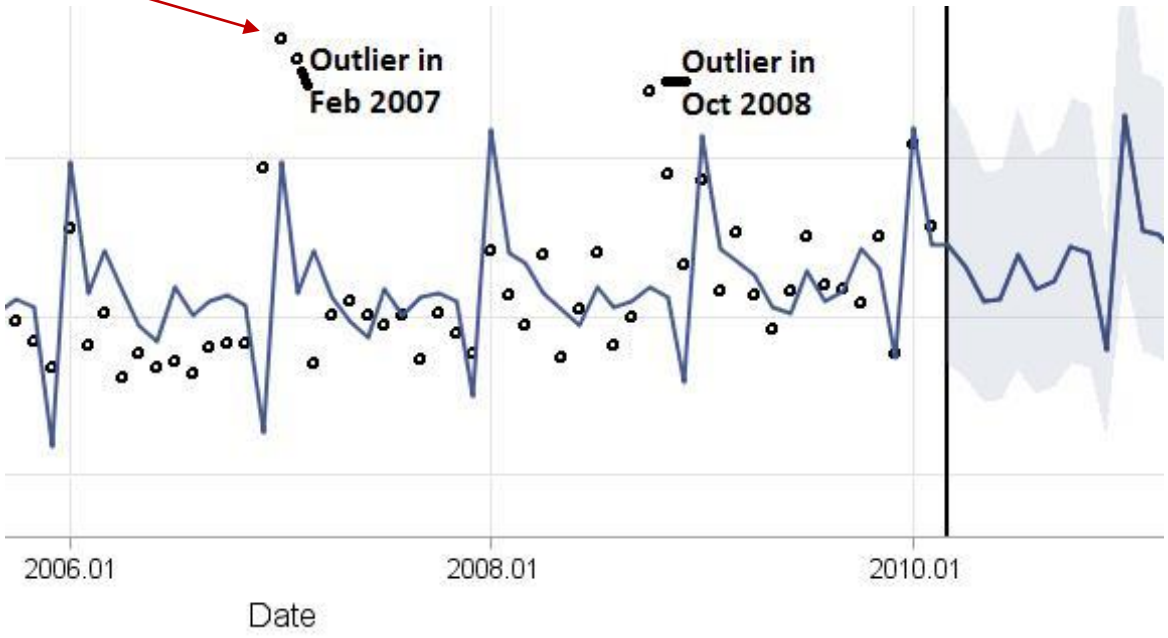




- Linear Regression
- Decision Trees
- Time Series Analysis

„Ja, aber …. im Jänner haben wir immer deutlich mehr Ereignisse“

Plausibler Wert für Jänner 2007
weil Jänner-Werte immer
höher sind

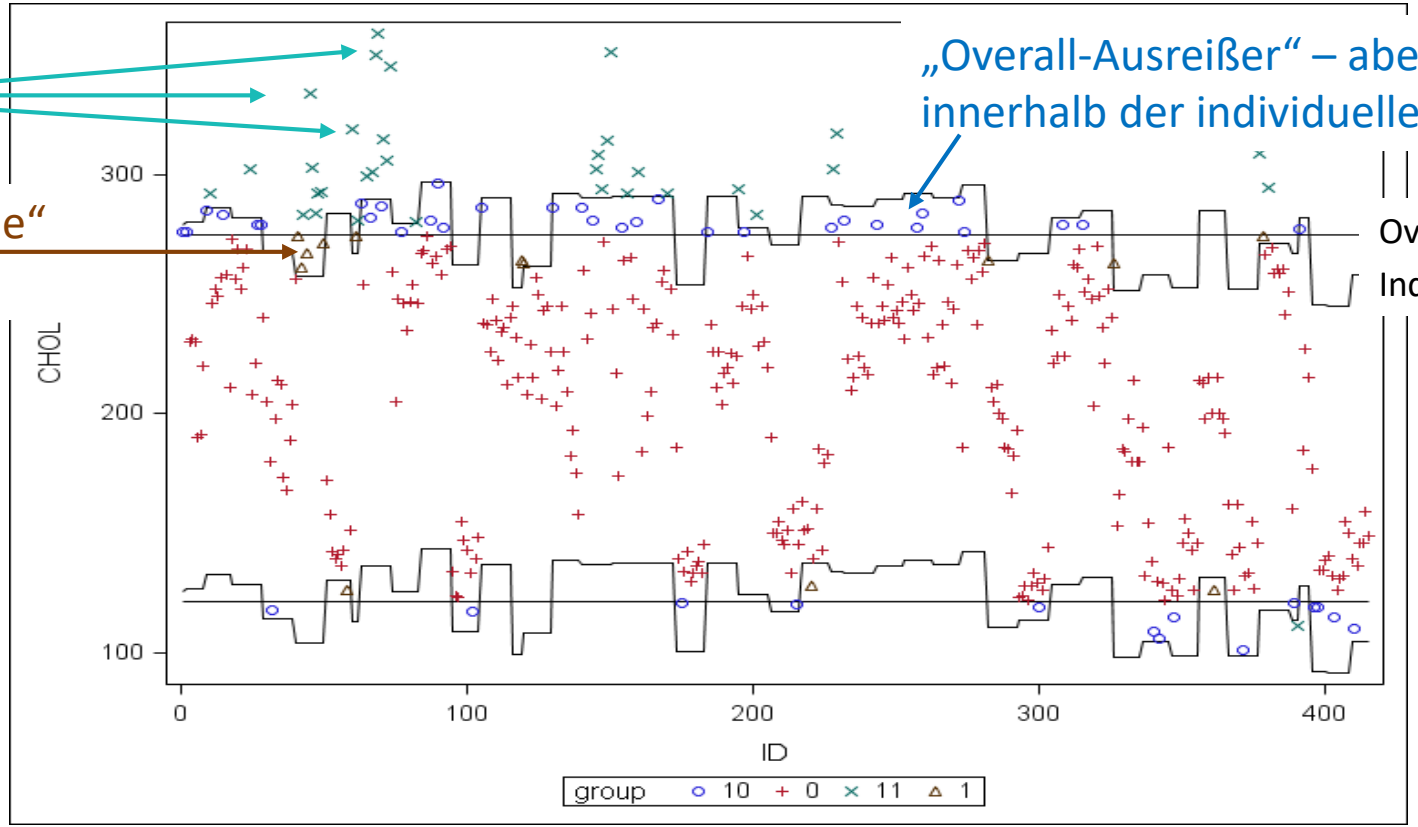


„Alle deren Wert größer x ist, sind Ausreißer! - Wirklich?“

Ausreißer

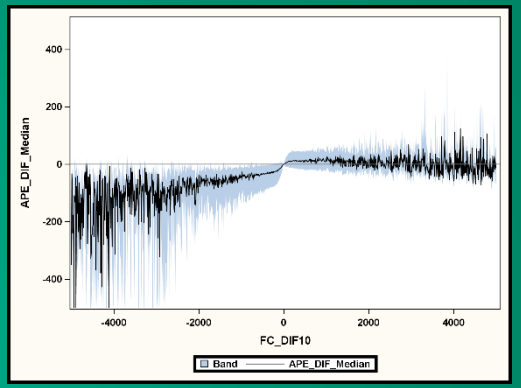
„Übersehene“ Ausreißer

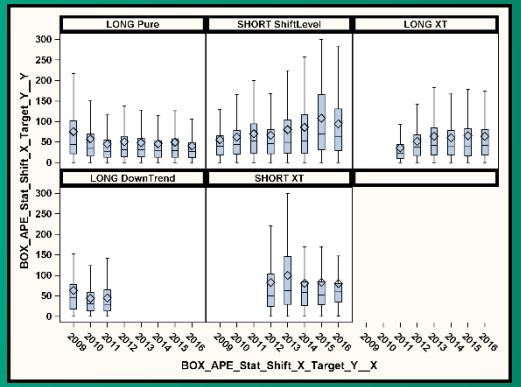
„Overall-Ausreißer“ – aber innerhalb der individuellen Limits



Data Science in Action: #4

Explaining Forecast Errors and Deviations



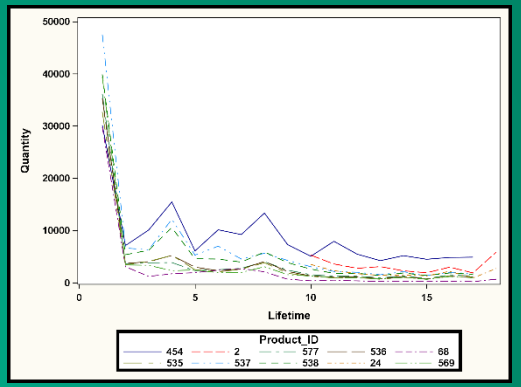


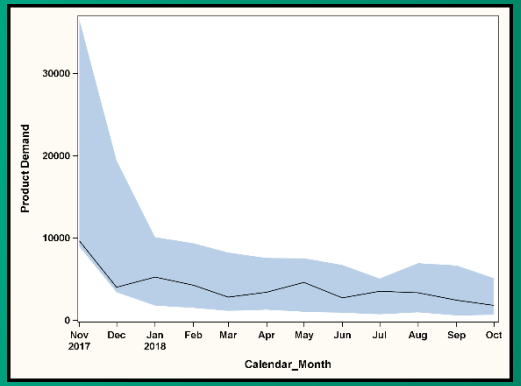
Do the demand planners really improve forecast accuracy with their manual overwrites?

Linear Regression
Quantile Regression
Descriptive Statistics

Data Science in Action: #5

Forecasting the Demand for New Products





Can the expected demand of products that are introduced only right now be estimated for forecast planning?

Poisson Regression
Cluster Analysis
Similarity Search

NOVELTY FORECASTING

- Training data from previous collections
- Generalized linear model

- Predictors
  - Product attributes
  - Time-dependant influence factors
  - Number of shops
  - Actual order intake
  - Actual sell-through

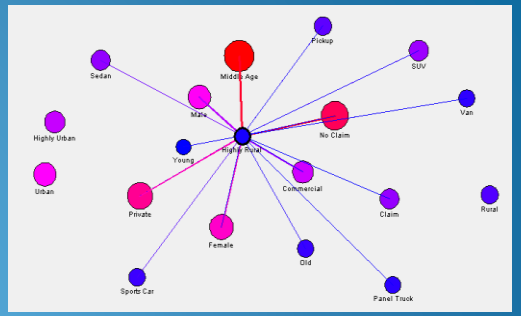
Nirvana ring

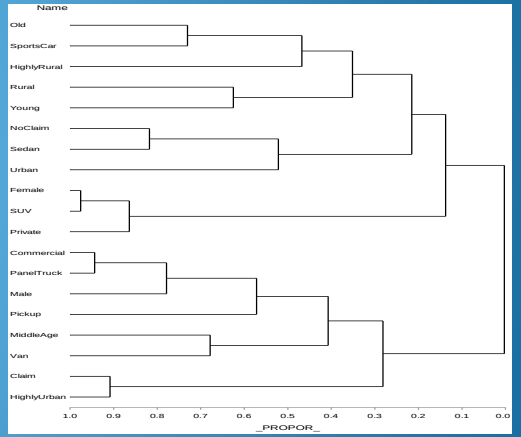
SWAROVSKI

Data Science in Action: #6

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

Can your data tell you stories about your analysis subjects, even if you don't ask explicitly?





Unsupervised machine learning methods:
association analysis
variable clustering

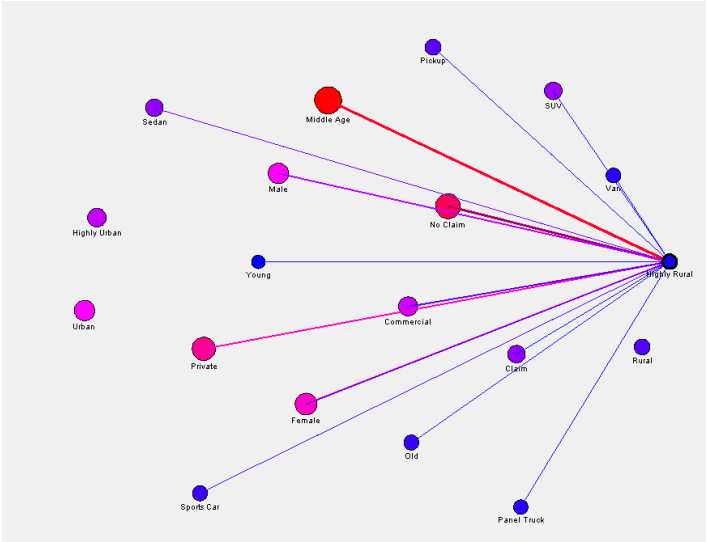
Lassen Sie ihre Daten sprechen!

Auffinden von Zusammenhängen in Ihren Analysedaten

- Daten aus der KFZ-Versicherung mit 6 Eigenschaften pro Versicherungsnehmer

Variable	Feature
AGE	YOUNG, MIDLIFE, OLD
GENDER	MALE, FEMALE
DENSITY	HIGHLY URBAN, URBAN, HIGHLY RURAL, RURAL
CAR_TYPE	VAN, SPORTS CAR, SUV, SEDAN, PICK UP
CAR_USAGE	PRIVATE, COMMERICIAL
CLM_FLAG	CLAIM, NO CLAIM

- Anwenden von unsupervised machine learning (Assoziationsanalyse) um Zusammenhänge zwischen den Eigenschaften aufzudecken.



Trauen Sie sich! Transponieren Sie die Daten, so wie Sie es sonst typischerweise nicht tun.

One-Row-Per-Subject

POLICYNO	CLM_FLAG	CAR_USE	CAR_TYPE	AGE	GENDER	DENSITY
160	No	Private	Sedan	60	M	Highly Urban
24836	No	Commercial	Sedan	43	M	Highly Urban
28046	No	Private	Van	48	M	Urban
28960	No	Private	SUV	35	F	Highly Urban
40933	No	Private	Sedan	51	M	Highly Urban
55277	No	Private	SUV	50	F	Urban
63212	Yes	Commercial	Sports Car	34	F	Highly Urban
69651	No	Private	SUV	54	F	Highly Urban
88070	Yes	Private	Sedan	40	M	Urban
93553	No	Commercial	SUV	44	F	Rural
127444	Yes	Commercial	Van	37	M	Highly Urban
141509	Yes	Private	SUV	34	F	Highly Urban
145326	No	Commercial	Van	50	M	Rural
146809	Yes	Private	Sports Car	53	F	Urban
148250	No	Private	Sedan	43	F	Rural
157851	No	Commercial	Van	55	M	Urban

Multple-Row-Per-Subject
Key-Value Tabelle

POLICYNO	Feature
160	Highly Urban
160	No Claim
160	Sedan
160	Private
160	Male
160	Old
24836	Highly Urban
24836	No Claim
24836	Sedan
24836	Commercial
24836	Male
24836	Middle Age

Lassen Sie ihre Daten sprechen!

Männer fahren kaum Sportwägen?

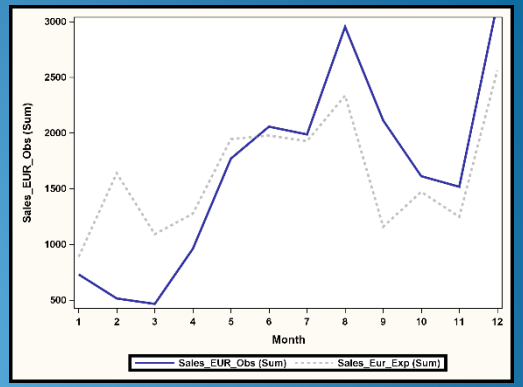
Regel 278 besagt, dass Sportwägen nur in 2,54 % der Fälle von Männern gefahren werden (erwartet wären 46 %)

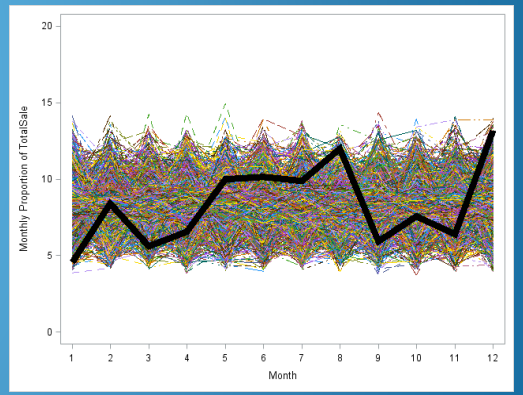
index	RULE	_LHAND	_RHAND	COUNT	SUPPORT	EXP_CONF	CONF	LIFT
267	Commercial ==> Sports Car	Commercial	Sports Car	200.00	1.94	11.44	5.28	0.46
268	Rural ==> Claim	Rural	Claim	102.00	0.99	26.66	6.52	0.24
269	Claim ==> Rural	Claim	Rural	102.00	0.99	15.18	3.71	0.24
270	Young ==> Highly Urban	Young	Highly Urban	10.00	0.10	34.93	8.33	0.24
271	Highly Rural ==> Claim	Highly Rural	Claim	32.00	0.31	26.66	6.30	0.24
272	Claim ==> Highly Rural	Claim	Highly Rural	32.00	0.31	4.93	1.17	0.24
273	Van ==> Female	Van	Female	117.00	1.14	53.82	12.70	0.24
274	Female ==> Van	Female	Van	117.00	1.14	8.94	2.11	0.24
275	Panel Truck ==> Female	Panel Truck	Female	40.00	0.39	53.82	4.69	0.09
276	Male ==> SUV	Male	SUV	99.00	0.96	27.98	2.08	0.07
277	SUV ==> Male	SUV	Male	99.00	0.96	46.18	3.43	0.07
278	Sports Car ==> Male	Sports Car	Male	30.00	0.29	46.18	2.54	0.06

- Kann anzeigen, dass in unserer Datenbasis tatsächlich Sportwägen in erster Linie von Frauen gefahren warden.
- Möglicherweise bietet ein Mitbewerber eine Polizze für Männer zu einem deutlich besseren Preis an.
- Ein fachliche Erklärung kann sein, dass der Sportwagen das 2. oder 3. Auto in der Familie ist, und dieser aus steuerlichen Gründe auf die Ehefrau registriert ist.
- Kann auch ein Trigger für eine detailliertere Analyse der Datenqualität sein.

Data Science in Action: #7

Checking the Alignment with Predefined Pattern





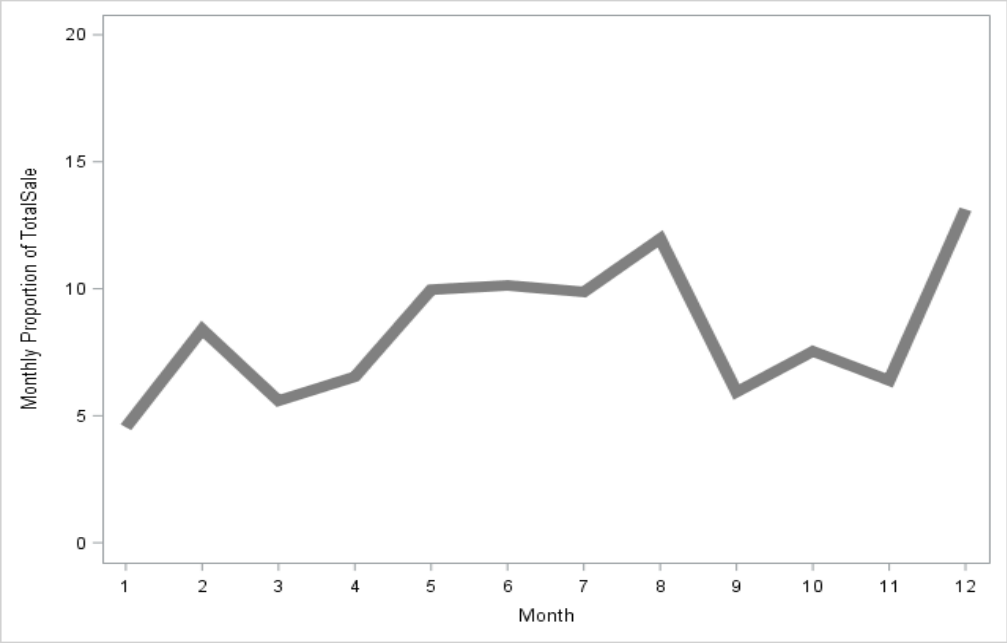
Which customers show a behavior that is far from what you expected?

Chi2 independency test
Benford's law
Time Series Similarity

“Welche meiner Verkäufer halten sich kaum an unsere Vorgaben?”

Der Bedarf an “Sub-Contracts” für ein Cateringunter-nehmen variiert im Verlauf eines Kalenderjahres

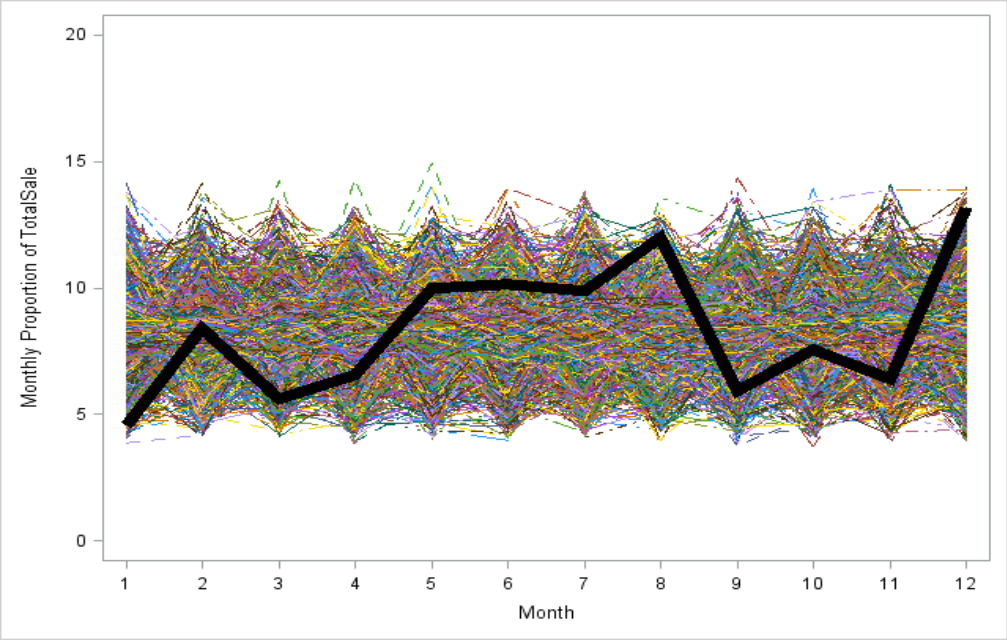
Verkäufer sind angehalten, entsprechend dieses Musters Verträge zu akquirieren.



Anzeige der Jahresverläufe pro Verkäufer hilft nicht wirklich

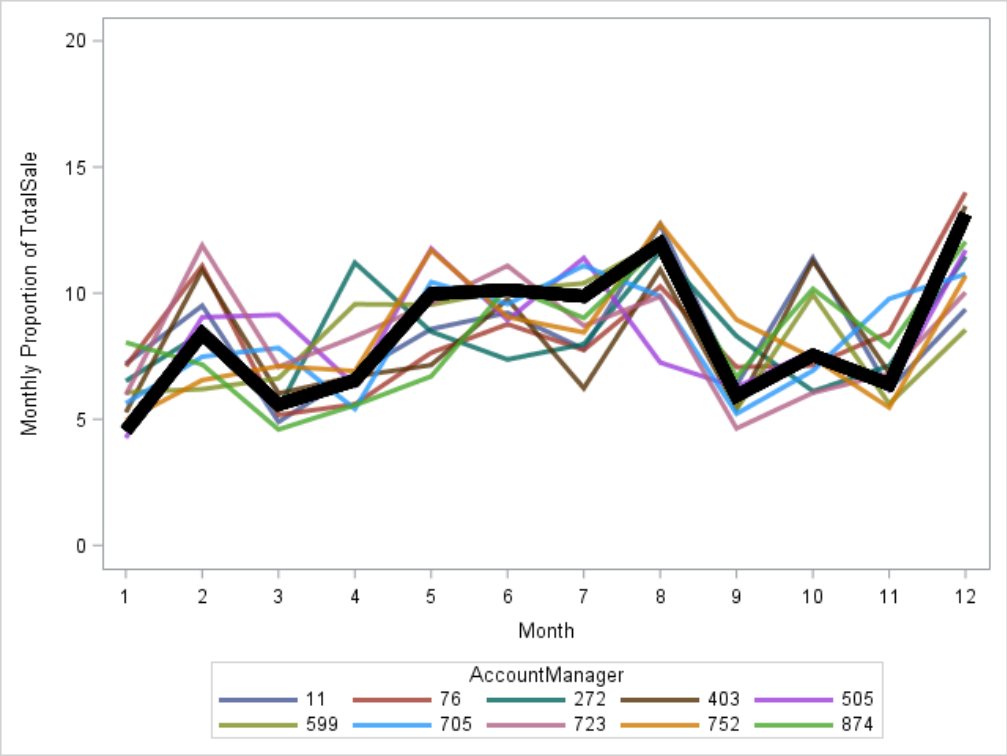
Kein klares Bild.

Unmöglich,
alle Linien einzeln
durchzusehen.



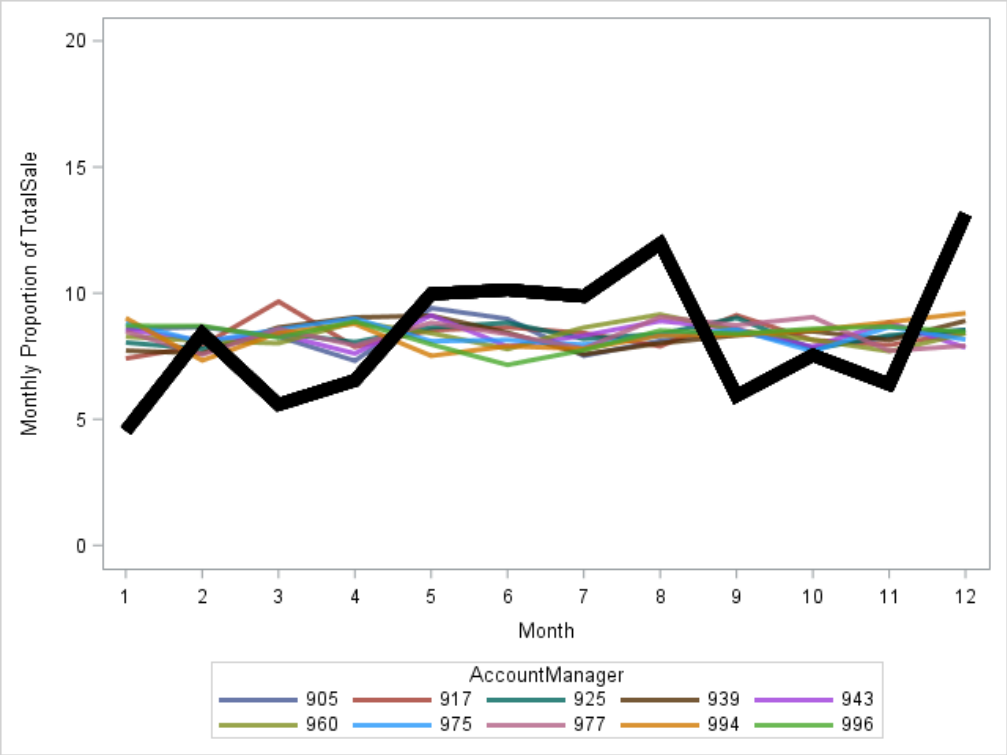
Ranking der Verkäufer mit analytischen Methoden (1)

Top 10 Verkäufer bzgl. “Alignment” mit der Vorgabe



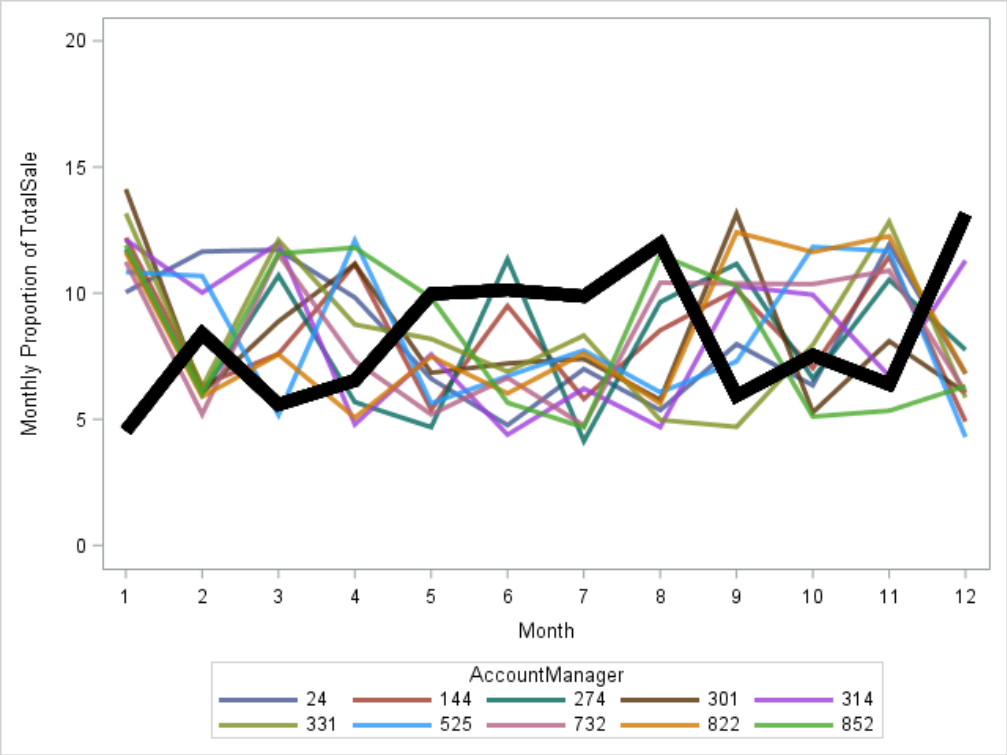
Ranking der Verkäufer mit analytischen Methoden (2)

Top 10 Verkäufer, für die es keine saisonale Variation gibt.

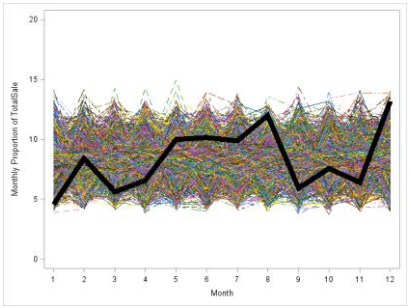


Ranking der Verkäufer mit analytischen Methoden (3)

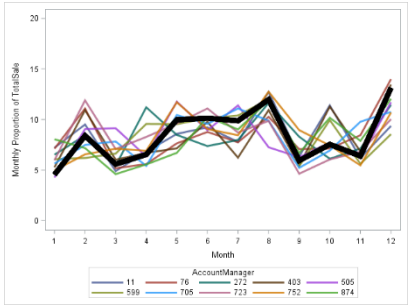
Top 10 Verkäufer die “gegen” das Muster arbeiten

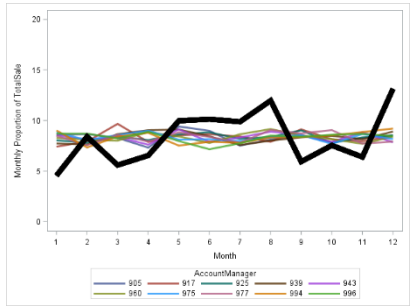


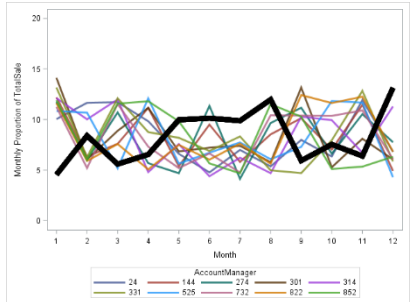
Analytik hilft mir, ein klareres Bild zu gewinnen!



Data Science Methoden





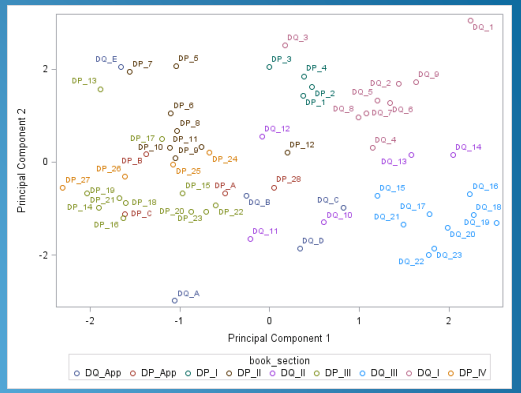


Vom „Rauschen“

zu interpretierbaren Segmenten

Data Science in Action: #8

Topic Search Documents and Clustering



Can I automatically find clusters of documents with similar content?

Text Mining
Text Parsing (Synonyme, Stemming, Stop-Listen)
Term by Document Weights

Kann ich ähnliche Kapitel erkennen, ohne die Bücher (von Gerhard ☺) erst lesen zu müssen?

Topic > +access,+file,+text,+relational,+relational database

PAGE 104 Data Preparation for Analytics Using SAS Chapter 13: Accessing Data PAGE 103 Part 3 Data Mart Coding and Content Chapter 13 Acces
Transposing One- and Multiple-Rows-per-Subject Data Structures 115 Chapter 15 Transposing Longitudinal Data 131 Chapter 16 Transformations of
Chapter 17 Transformations of Categorical Variables 161 Chapter 18 Multiple Interval-Scaled Observations per Subject 179 Chapter 19 Multiple Catego

PAGE 38 Data Preparation for Analytics Using SAS Chapter 5: The Origin of Data PAGE 43 Part 2 Data Structures and Data Modeling Chapter 5 The
Models 45 Chapter 7 Analysis Subjects and Multiple Observations 51 Chapter 8 The One-Row-per-Subject Data Mart 61 Chapter 9 The Multiple-Rows-p
Data Structures for Longitudinal Analysis 77 Chapter 11 Considerations for Data Marts 89 Chapter 12 Considerations for Predictive Modeling 95 Introdu

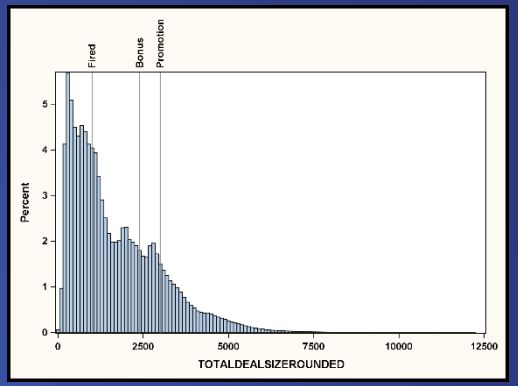
PAGE 178 Data Preparation for Analytics Using SAS Chapter 17: Transformations of Categorical Variables PAGE 177 Chapter 17 Transformations
Introduction 17.2 General Considerations for Categorical Variables 162 17.3 Derived Variables 164 17.4 Combining Categories 166 17.5 Dummy Codi
Multidimensional Categorical Variables 172 17.7 Lookup Tables and External Data 176 17.1 Introduction In this chapter we will deal with transformatio

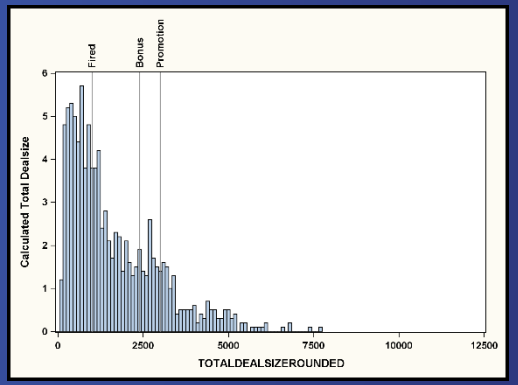
40 Data Quality for Analytics Using SAS Chapter 3: Data Availability 41 Chapter 3: Data Availability 3.1 Introduction 32 3.2 General Considerations 32 Re
data availability 32 Availability and usability 32 Effort to make data available 33 Dependence on the operational process 33 Availability and alignment in t
of Historic Data 34 Categorization and examples of historic data 34 The length of the history 35 Customer event histories 35 Operational systems and a

PAGE 382 Data Preparation for Analytics Using SAS Appendix B: The Power of SAS for Analytic Data Preparation PAGE 381 Appendix B The Power
369B.1 Motivation B.2 Overview 370 B.3 Extracting Data from Source Systems 371 B.4 Changing the Data Mart Structure: Transposing 371 B.5 Data Mar
Multiple-Rows-per-Subject Data Sets 372 B.6 Selected Features of the SAS Language for Data Management 375 B.7 Benefits of the SAS Macro Langu

Data Science in Action: #9

Using Monte Carlo Simulations to Understand the Outcome Distribution





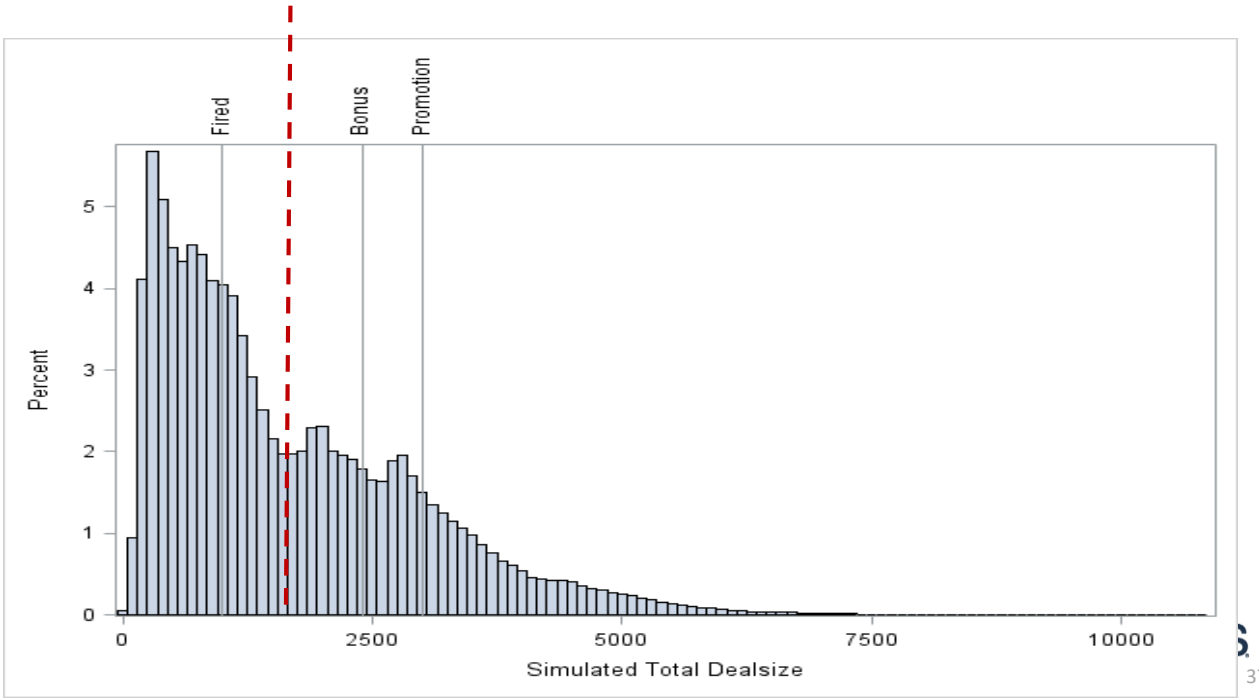
When the sales manager looks at the project pipeline, does the sum of weighted averages give him or her a full picture?

Monte Carlo simulations
Mathematical programming

Wird der Sales Manager seinen Job behalten?

ProjectID	DealSize (1000 $)	Proba-bility
1	1500	10%
2	10	65%
3	500	20%
4	50	50%
5	100	40%
6	30	90%
7	10	60%
8	150	20%
9	200	25%
10	180	10%
11	900	10%
12	750	20%
13	600	10%
14	320	20%
15	100	40%
16	50	80%
17	2000	5%
18	400	20%
19	2500	10%
20	1700	15%

Gewichtetes Mittel:
$ 1.661.500

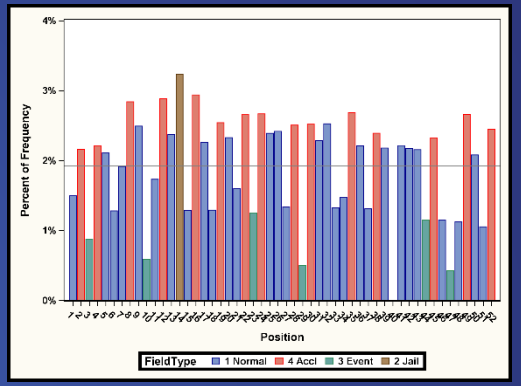


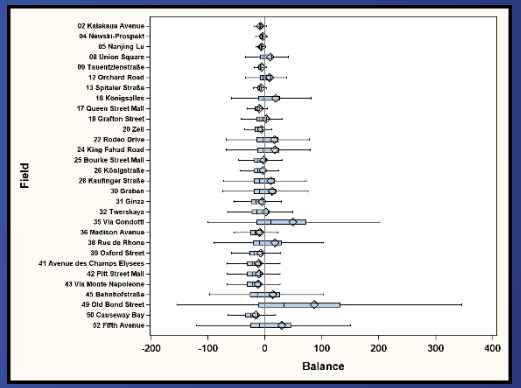
Data Science
in Action: #10

Studying Complex Systems – Simulating the Monopoly Board Game

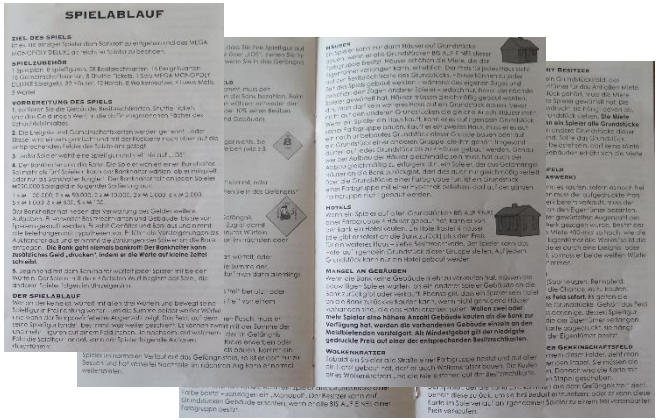
How can you simulate complex environments to get insight in the most frequent processes?

Monte Carlo Simulations

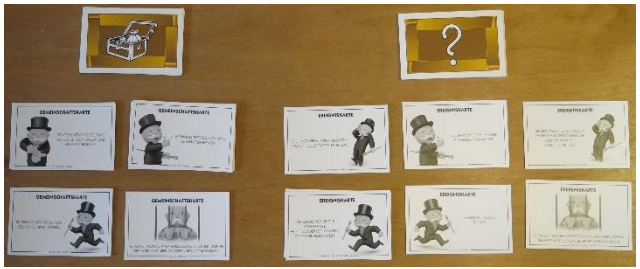




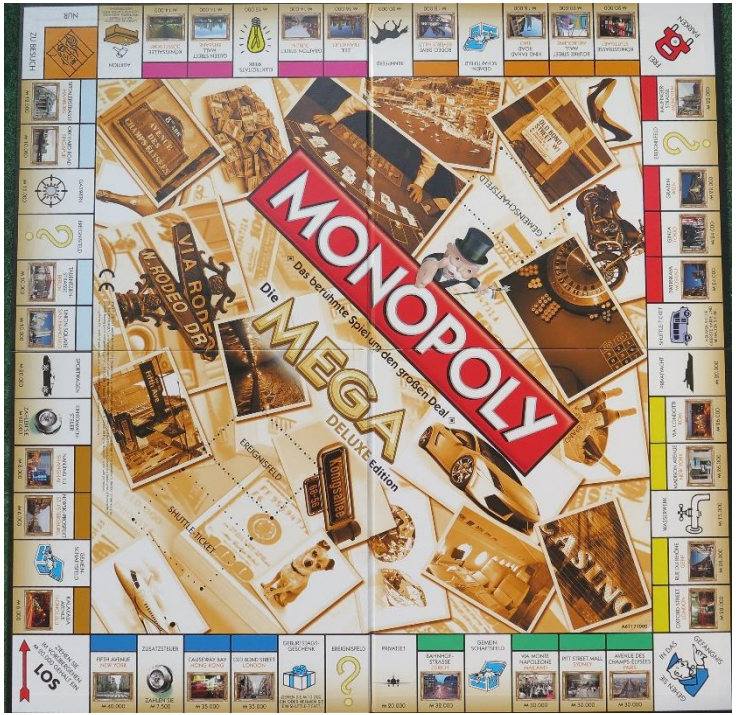
Das Monopoly Spiel ist vielen Frameworks im Geschäftsleben gar nicht so unähnlich



Komplexe Regeln



Zusätzliche Anweisungen



Rahmenwerk von Möglichkeiten und Ereignissen

Monetäre Dimension

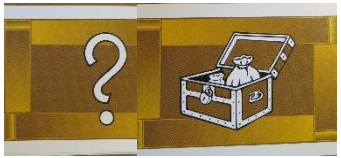
Dynamische Komponenten

Zufällige Komponenten

Simulation komplexer Prozesse erlaubt mir Einblick in Zusammenhänge (die ich sonst nicht gesehen hätte)

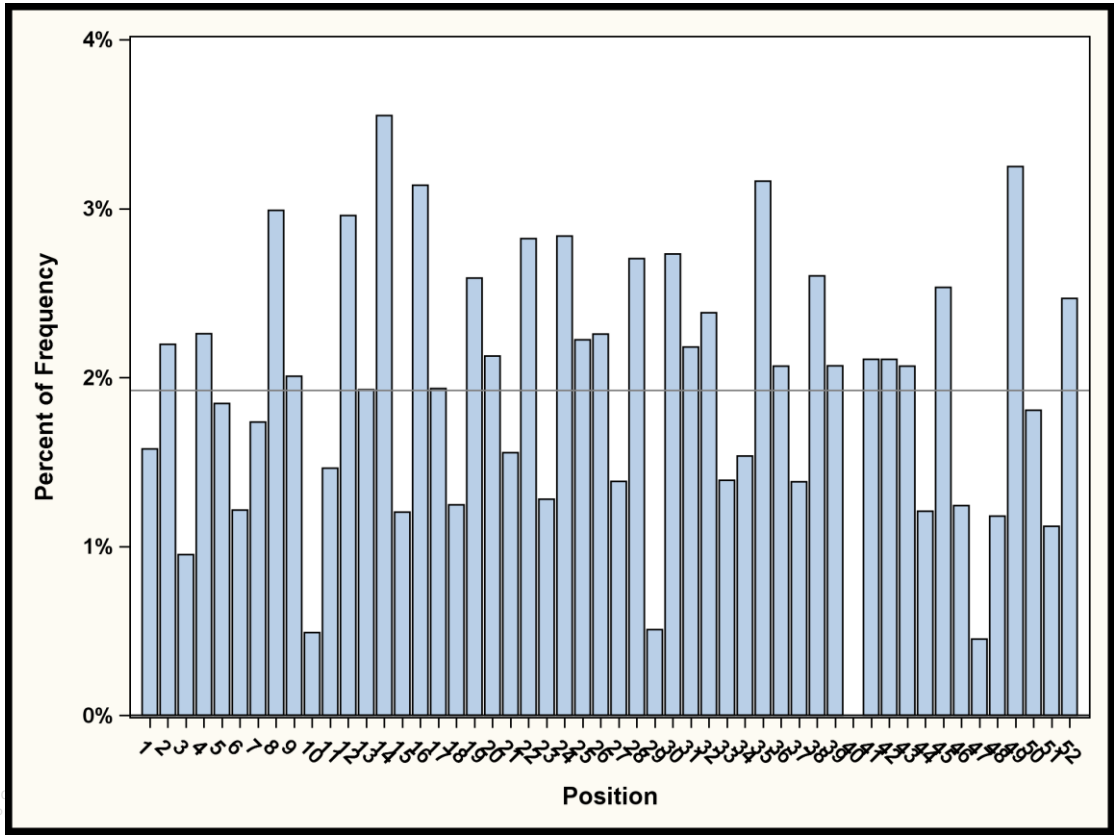
Würfel-Summe

Gehe ins Gefängnis!

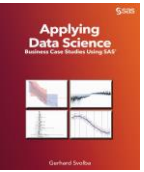


Ereignis Felder

Accelerator Würfel



Further Links and Downloads



- Cases #1-2, 4-7, 9-10:
  - http://www.sascommunity.org/wiki/Applying_Data_Science_-_Business_Case_Studies_Using_SAS
  - http://www.sascommunity.org/wiki/DOWNLOAD_SECTION:_Applying_Data_Science_-_Business_Case_Studies_Using_SAS
- #1 – Survival
  - *SAS/STAT® 14.2 User's Guide. The LIFETEST Procedure*. http://support.sas.com/documentation/onlinedoc/stat/142/lifetest.pdf (accessed 1 March 2017).
  - Allison, P. 1995. *Survival Analysis Using SAS®: A Practical Guide, Second Edition*. Cary, NC: SAS Institute Inc.
- #2 – Detecting Breakpoints and Outliers
  - Kuhfeld, W., and W. Cai. 2013. "Introducing the New ADAPTIVEREG Procedure for Adaptive Regression." SAS Global Forum Proceedings. http://support.sas.com/resources/papers/proceedings13/457-2013.pdf (Paper 457-2013).
- #3 – Individual Reference Values: http://www.sascommunity.org/wiki/Data_Quality_for_Analytics
- #4 – Forecast Error Analysis
  - SGF2018 – Paper 1673 - *Getting More Insight into Your Forecast Errors with the GLMSELECT and QUANTSELECT Procedures*
  - KSFE 2015: Gerhard Svolba: Mehr als linear oder logistisch – ausgewählte Möglichkeiten neuer Regressionsmethoden in SAS - Download the presentation and the paper

Further Links and Downloads (Forts.)

- #6 – Feature Data Mining: http://www.sascommunity.org/wiki/Data_Preparation_for_Analytics
- #8 – Text Mining
  - KSFE 2017 : Beitrag „SAS Text Analytics findet Zusammenhänge in Texten – Ergebnisse eines Selbstversuchs“
  - SAS Club 2015: SAS Contextual Analysis in Action – Erfahrungen aus einem Selbstversuch
- #9 – Sales Manager Simulation
  - SAS Club : 2016, Mihai Paunescu: Simulationen und Mathematische Programmierung mit SAS
  - KSFE 2018 (to be prepared)
- #10 – Monopoly Simulation
  - KSFE 2017 : "Gewinnen beim Monopoly® Spiel – Alles nur Zufall? Oder gibt es doch ein paar Muster, die man kennen sollte?"
  - SAS Club 2007: Simulationen und Monte-Carlo Analysen mit SAS

Key Takeaways

Analytics und Data Science sind da um Ihnen zu helfen!

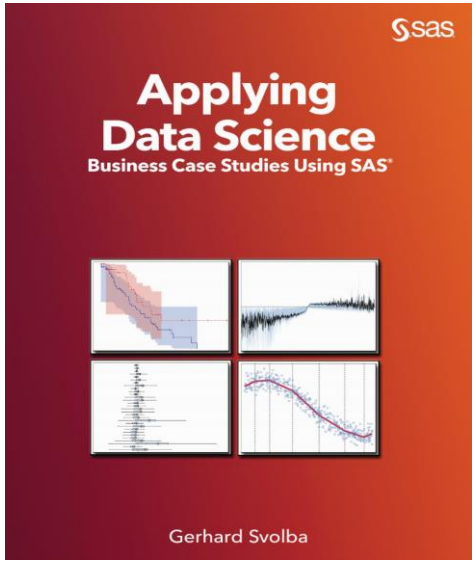
- Sie sehen ein klareres, objektiveres Bild Ihrer Daten und Analyse-Subjekte
- Sie erhalten explizite Ergebnisse anstatt die Nadel im Heuhaufen zu suchen
- Die Daten sprechen zu Ihnen und Sie erhalten die Ergebnisse automatisch statt manuell
- Do it again! – Behandeln Sie Ihre Modelle als “Asset” und wiederholen Sie Ihre Analyse

Machine Learning and Data Science sind das Kernstück der SAS Analytic Platform

- Umfassendes Set an Methoden – Entdecken und Produktivstellen
- Offen für unterschiedliche Benutzertypen (Coding, Point&Click, SAS, R, Python, …)

More Information

Gerhard Svolba – Principal Analytic Solutions Architect
sastools.by.gerhard@gmx.net
https://github.com/gerhard1050/



- Applying Data Science – Business Case Studies Using SAS, SAS Press 2017
- Eight Case Studies showing how Data Science and Analytics can be applied to provide insight into yout data and improve your business decisions
- http://www.sascommunity.org/wiki/Applying_Data_Science_-_Business_Case_Studies_Using_SAS