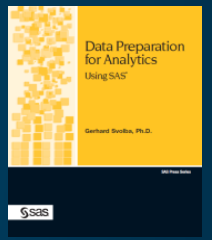
Kann ich die Verweildauer meiner Mitarbeiter analysieren und vorhersagen? Survival Analyse von SAS liefert die Antworten

Gerhard Svolba, SAS Austria

Mannheim, 2. März 2018 - KSFE 2018

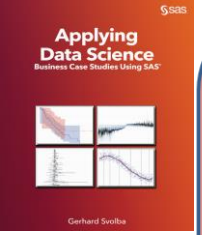


https://github.com/gerhard1050/

THE POWER TO KNOW.

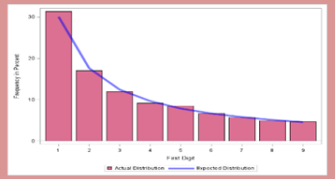
Überblick

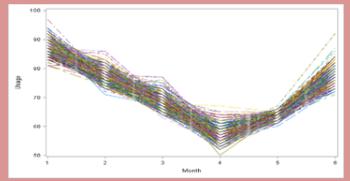
- Fachlicher Hintergrund des Fallbeispiels
- Problematik zensierter Daten
- Kaplan-Meier Methode und die LIFETEST Procedure
- Analyse von Einflussfaktoren mit der PHREG Procedure
- Ausgewählte Graphiken für die Mitarbeiter-Verweildauer
- Schlussfolgerungen



Checking the Alignment with Predefined Pattern

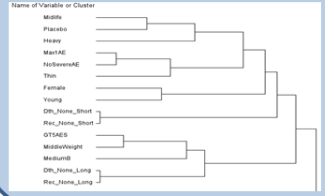
Which customers show a behaviour which is far from what you expected?

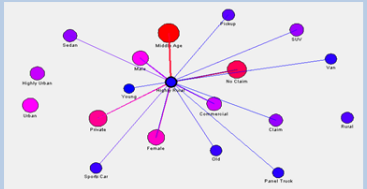




Listen to Your Data – Discover Unknown Relationships

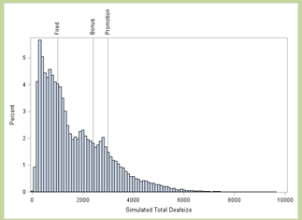
Can your data tell you stories, even if you don't ask them?

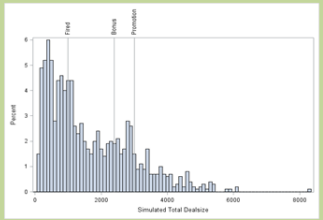




Using Monte Carlo Simulations to Understand the Outcome Distribution

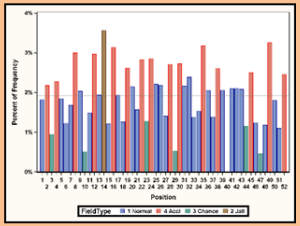
Will the Sales Manager keep his job (when you look at his sales pipeline)?

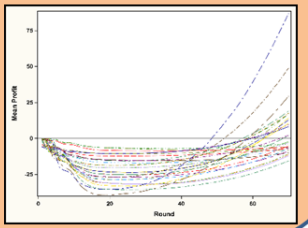


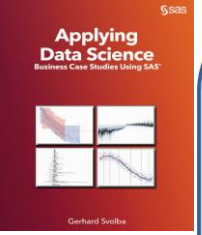


Studying Complex Systems – Simulate the Monopoly® Board Game

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

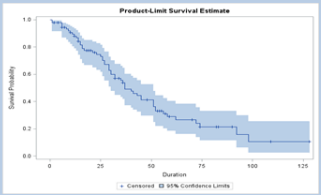


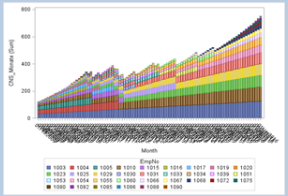




Performing Headcount Survival Analysis for Employee Retention

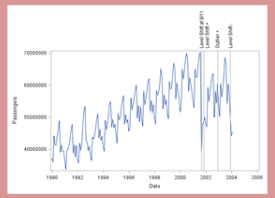
Can you make assumptions about the average length of time intervals, even if most of the endpoints have not yet been observed?

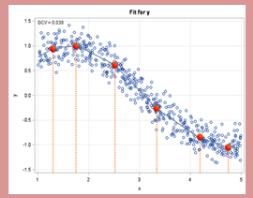




Detecting Outliers and Structural Changes in Longitudinal Data

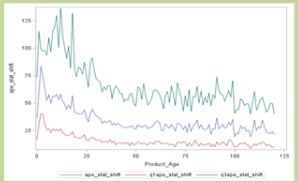
Can you automatically detect events and changes in the course of your data over time?

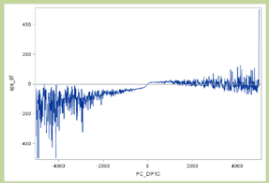




Explaining Deviations and Forecast Errors

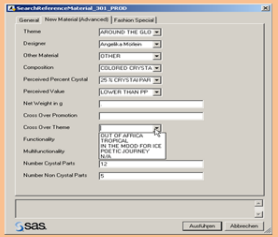
Do the demand planners really improve forecast accuracy with their manual overwrites?

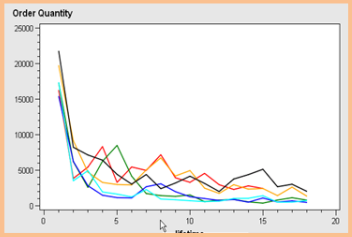




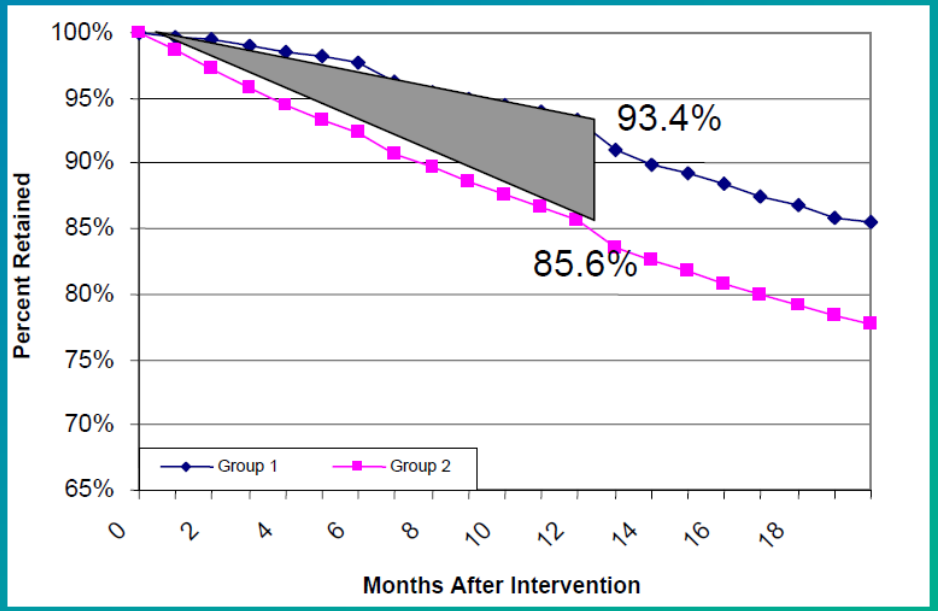
Forecasting the Demand for New Products

Can you assess the exptected demand of products that are introduced right now?





We Can Use Area to Quantify Results



- Increase in survival is given by the area between the curves.
- For the first year, area of triangle is a good enough estimate

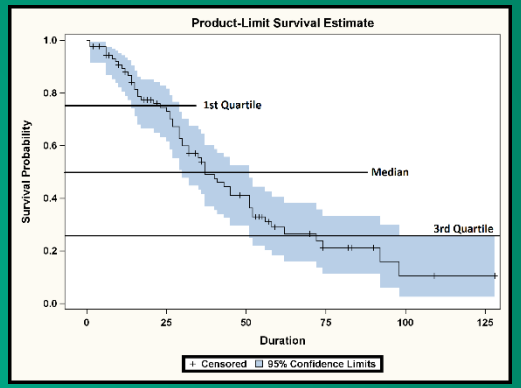
Note: there are easy ways to calculate the exact value

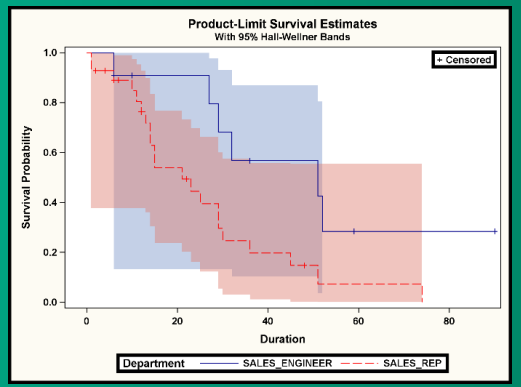
Data Miners

sas

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

Beispiel aus dem „Human Ressources“ Bereich

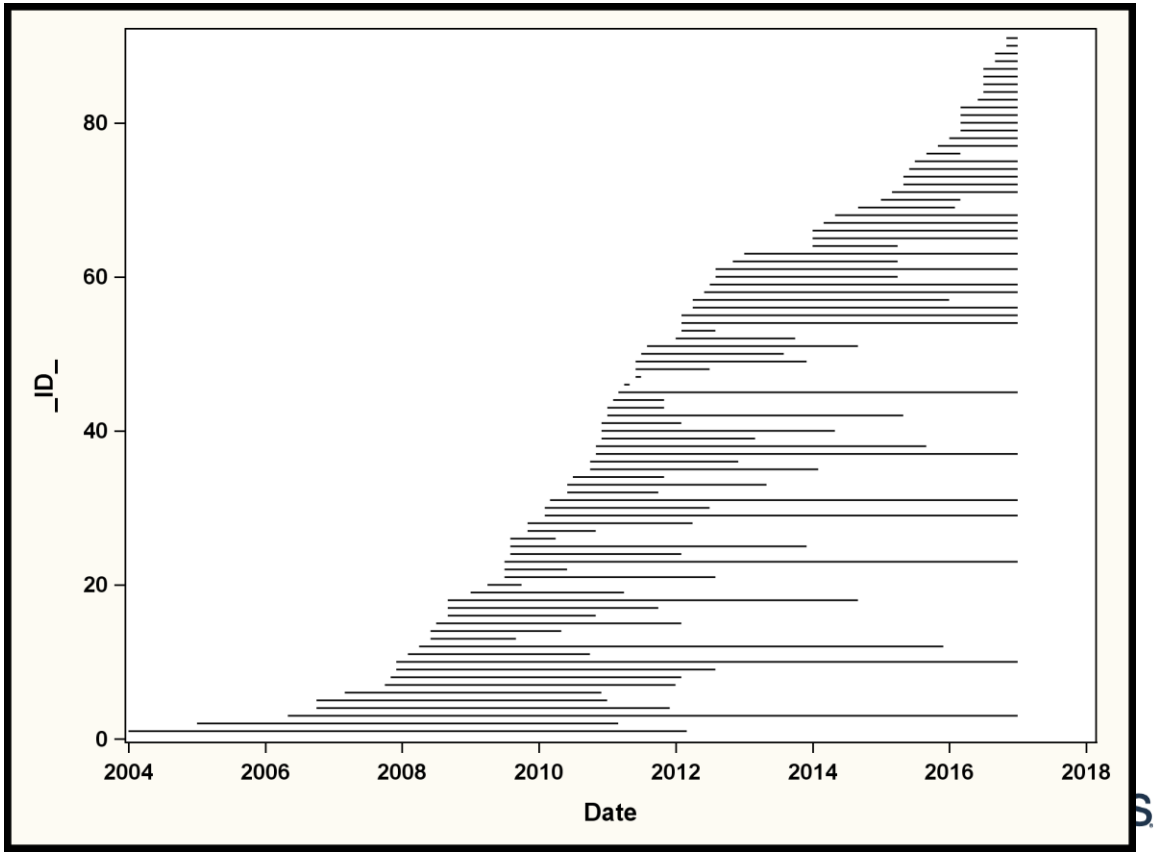
- Verweildauer von Mitarbeitern im Unternehmen
- Getrennt nach Abteilungen: Marketing, Admin, Sales, TechSupport, Sales Engineer

Variable Name
EmpNo
FirstName
Gender
Department
TechKnowHow
Start
End

EmpNo	FirstName	Department	Gender	Start	End	Status	Duration
1021	Mary	MARKETING	F	01JUL2009	01AUG2012	0	37
1022	Frank	SALES_REP	M	01JUL2009	01JUN2010	0	11
1023	Alan	SALES_ENGINEER	M	01JUL2009	.	1	90
1024	Frencesca	ADMINSTRATION	F	01AUG2009	01FEB2012	0	30
1025	Karl	SALES_ENGINEER	M	01AUG2009	01DEC2013	0	52
1026	Hana	ADMINSTRATION	F	01AUG2009	01APR2010	0	8
1027	Brian	SALES_REP	M	01NOV2009	01NOV2010	0	12
1028	Pawel	SALES_REP	M	01NOV2009	01APR2012	0	29
1029	Alessandro	TECH_SUPPORT	M	01FEB2010	.	0	83

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

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



Fachliche Fragen

- What is the average retention period for employees in the company?
- How can the important fact that the employment end date is known only for those who already left the company, be adequately considered in the analysis?

- How can the retention period be visualized and compared between different subgroups?

- Are there influential factors for the length of the retention period?
- How can these factors be ranked by magnitude of their influence?
- Can the expected survival period for an employee be predicted?

Ergebnisse der Kaplan-Meier Analyse

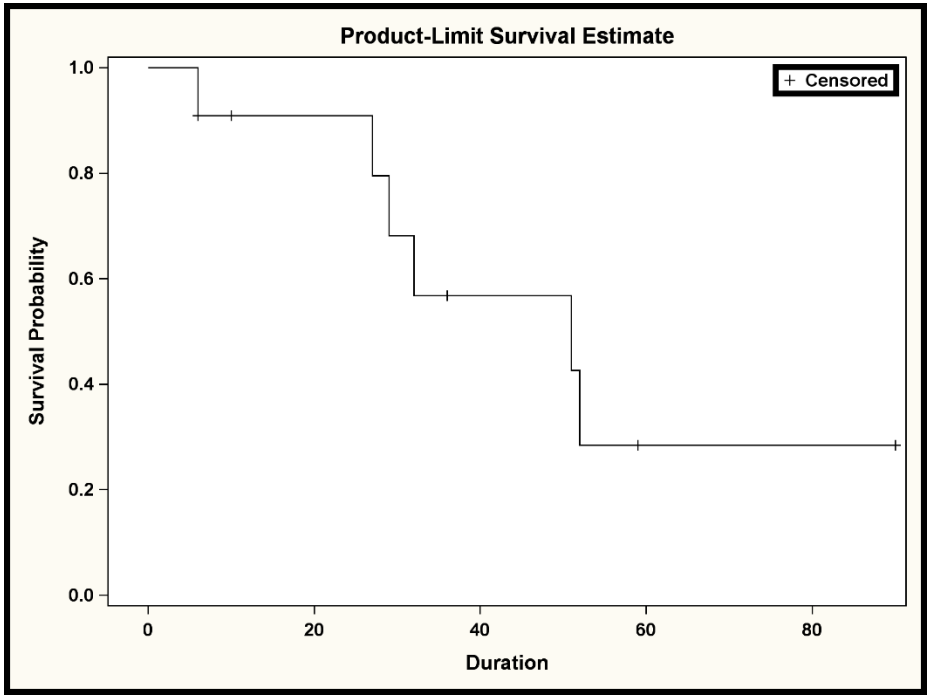
Sales-Engineer Department

Duration	Left	Resigned	Censored	Survival	Comment
0	11			1,000	Start of Observation
6	10	1	0	0,909	John resigns
6	9	0	1		Brady is censored from the analysis
10	8	0	1		Lucas is censored from the analysis
27	7	1	0	0,795	Rainer resigns
29	6	1	0	0,682	Vincenz resigns
32	5	1	0	0,568	George resigns
36	4	0	1		Mark is censored from the analysis
51	3	1	0	0,426	Viktor resigns
52	2	1	0	0,284	Karl resigns
59	1	0	1		Eugene is censored from the analysis
90	0	0	1	0,284	Alan is censored from the analysis

Kaplan-Meier Analyse mit der LIFETEST Procedure

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	32.0000	.
50	51.0000	LOGLOG	27.0000	.
25	29.0000	LOGLOG	6.0000	51.0000

Mean	Standard Error
39.9489	5.2333



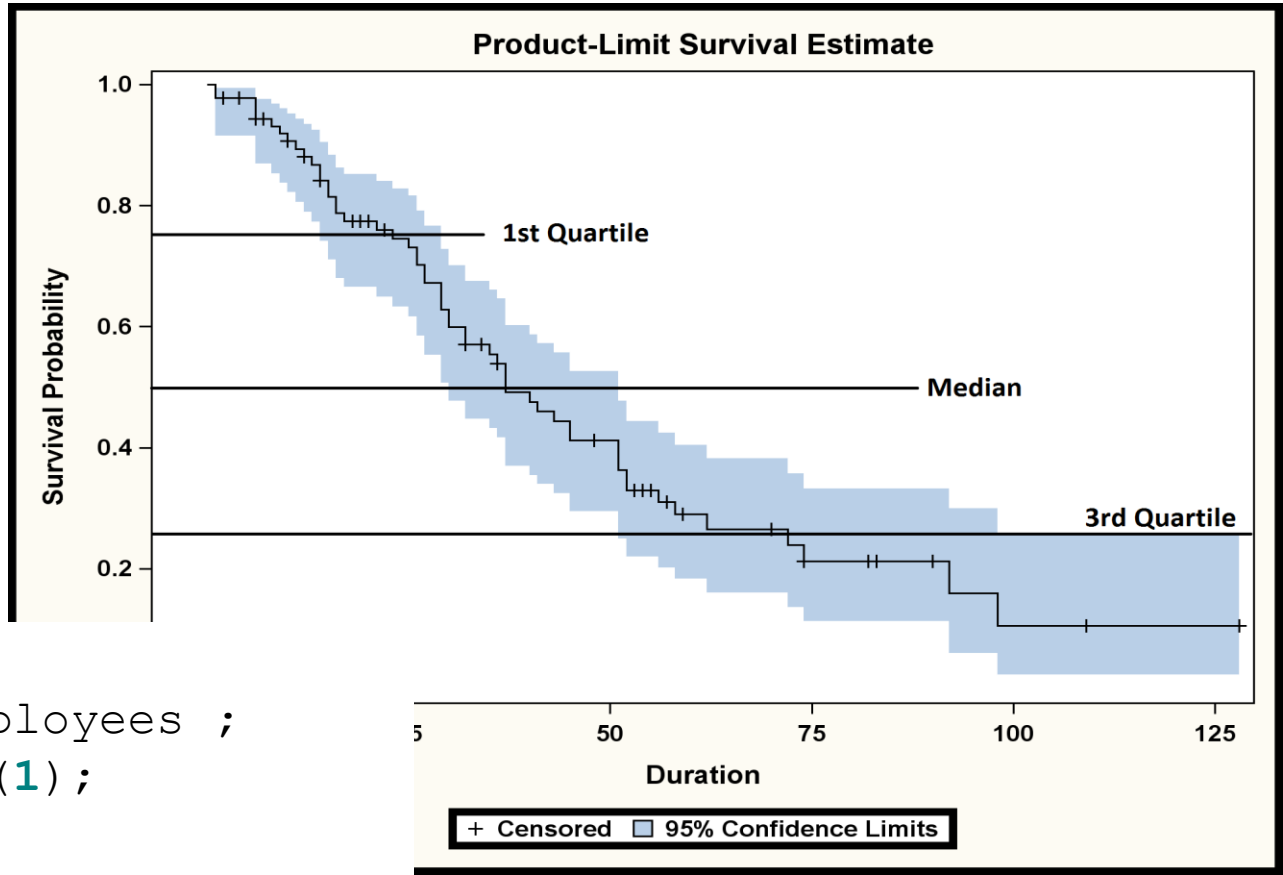

```
ods graphics on;
proc lifetest data=employees ;
 time Duration*Status(1);
 where Department='SALES_ENGINEER';
run;
```

Interpretation der Survival Kurve

Für alle Abteilungen

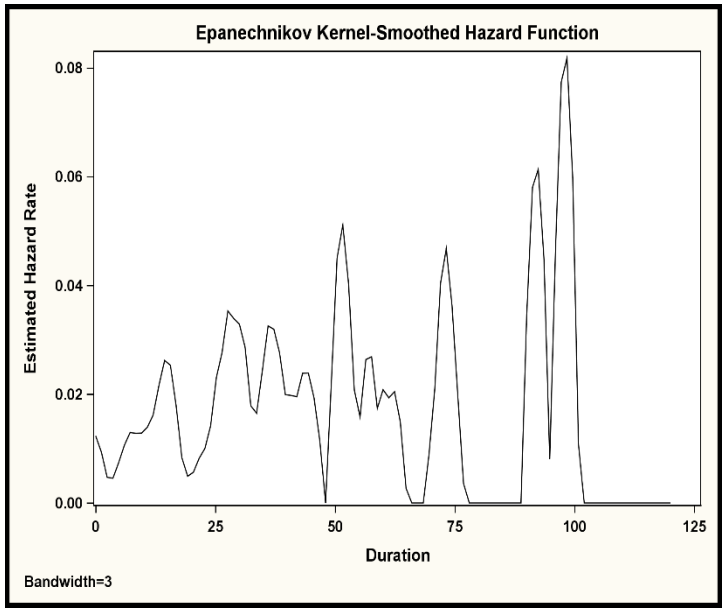
Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	72.000	LOGLOG	51.00	.
50	37.000	LOGLOG	30.00	51.00
25	23.000	LOGLOG	14.00	29.00

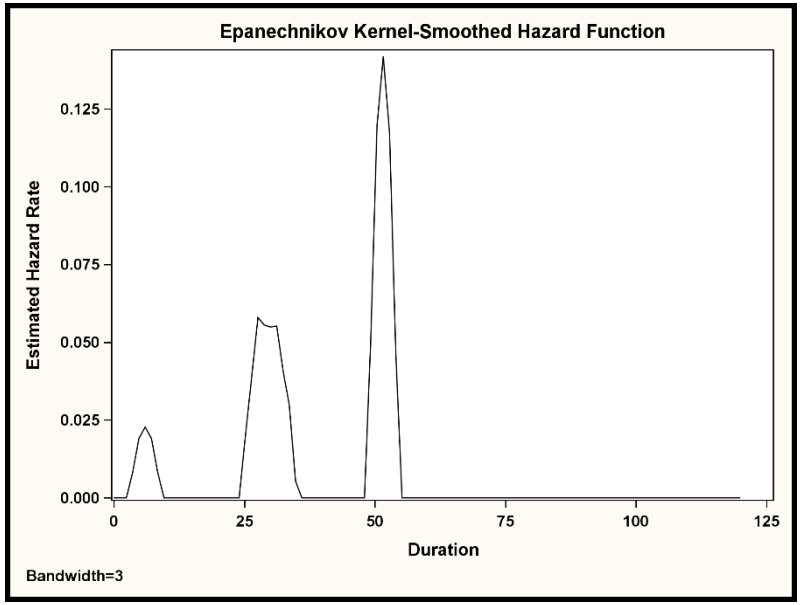
Mean	Standard Error
46.757	3.813




```
ods graphics on;
proc lifetest data=employees ;
 time Duration*Status(1);
run;
```

Analyse der Hazard Kurve

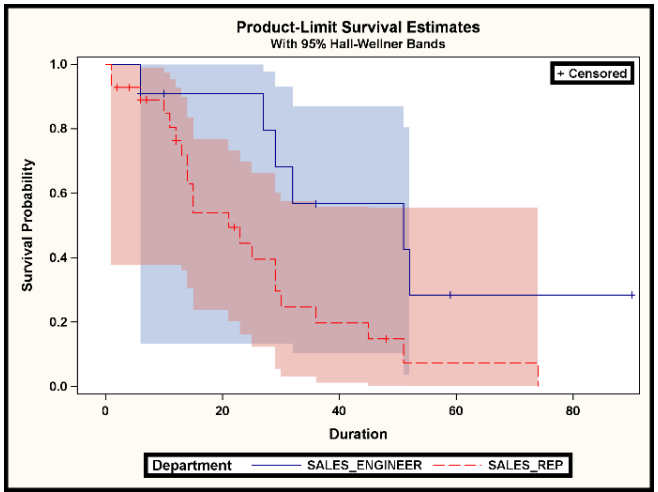


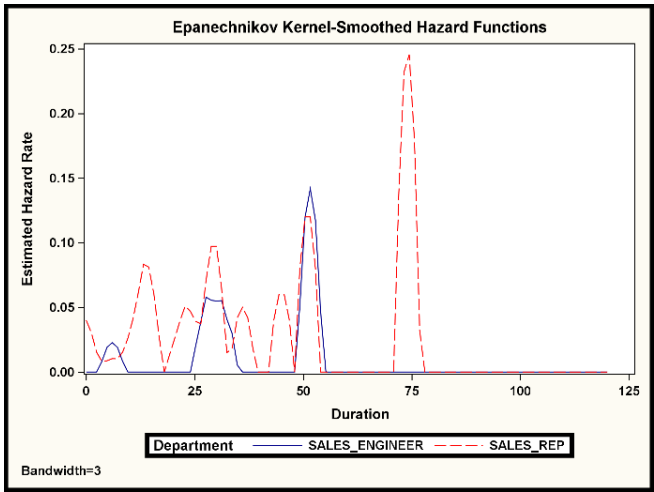



```
PROC LIFETEST DATA=employees plots=(hazard(bandwidth=3 maxtime=120));
 TIME Duration*Status(1);
RUN;
```

Konfidenz-Bänder mit der LIFETEST Procedure

```
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;
```

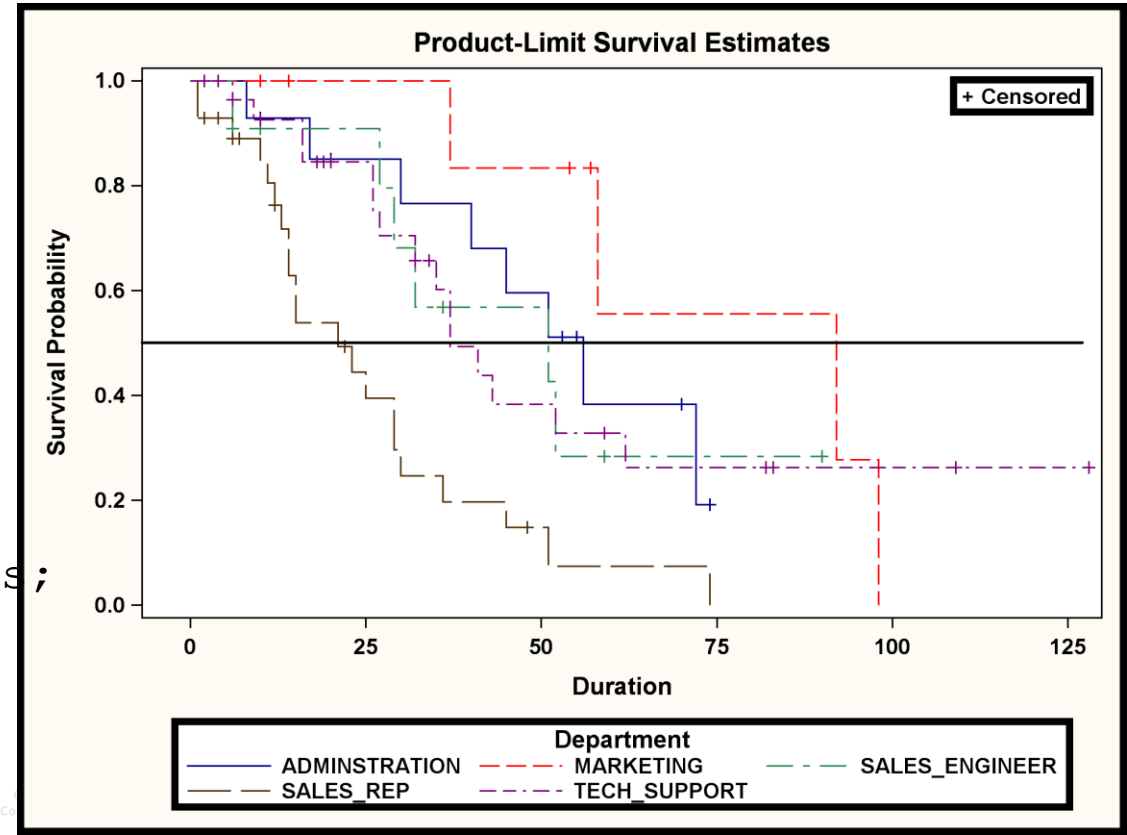




Survival-Kurve pro Abteilung

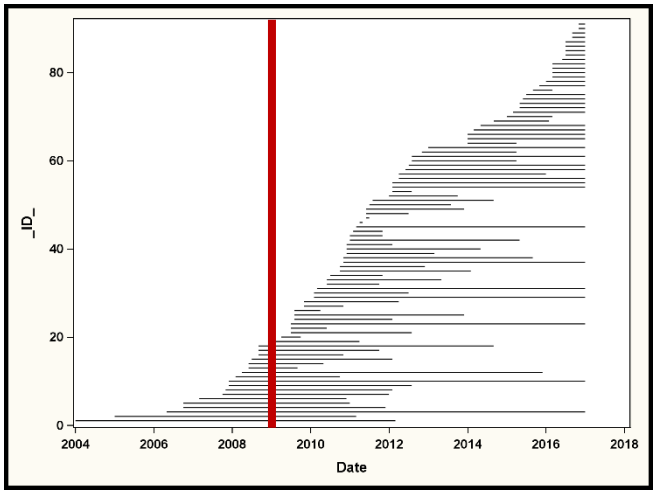
Referenz-Linie für den Median

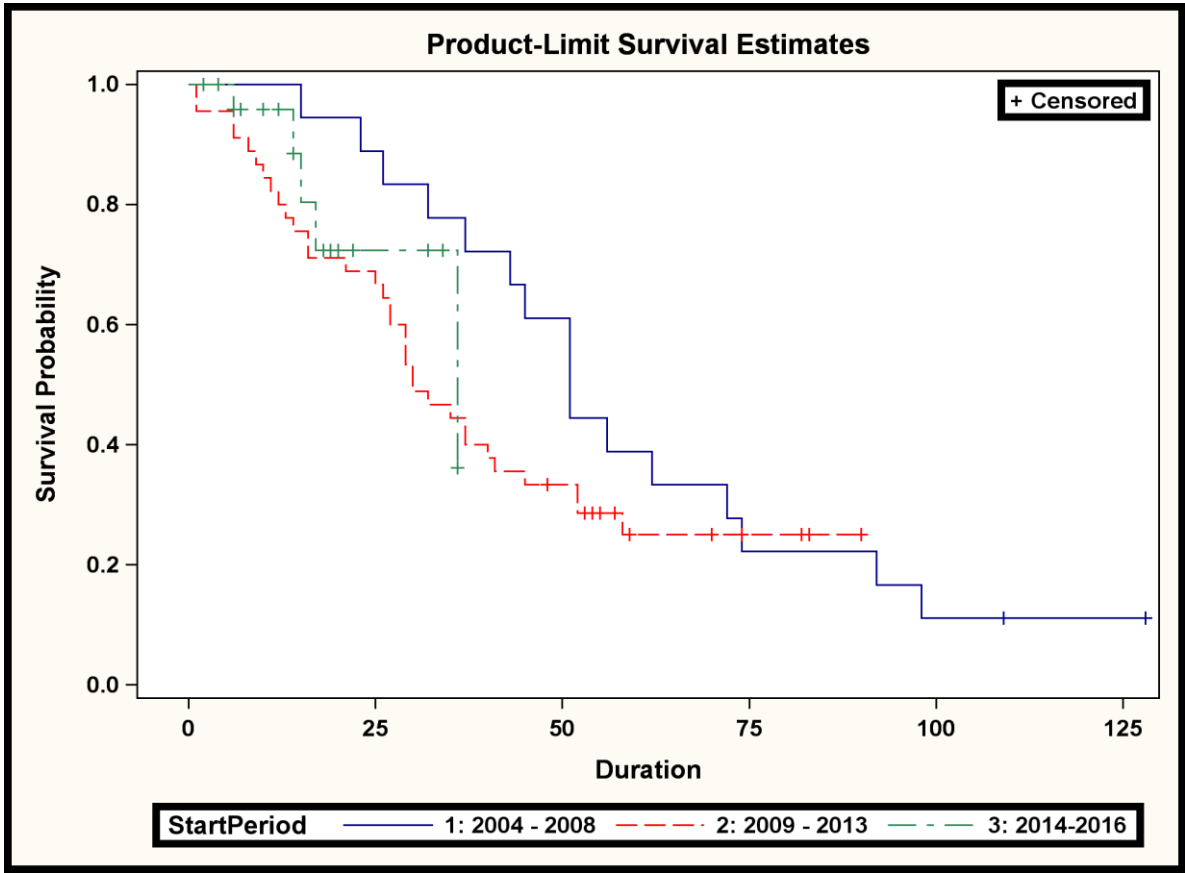
```
PROC LIFETEST DATA=employees;
  TIME Duration*Status(1);
  STRATA department;
RUN;
```



In den „guten alten Zeiten“ war alles beser

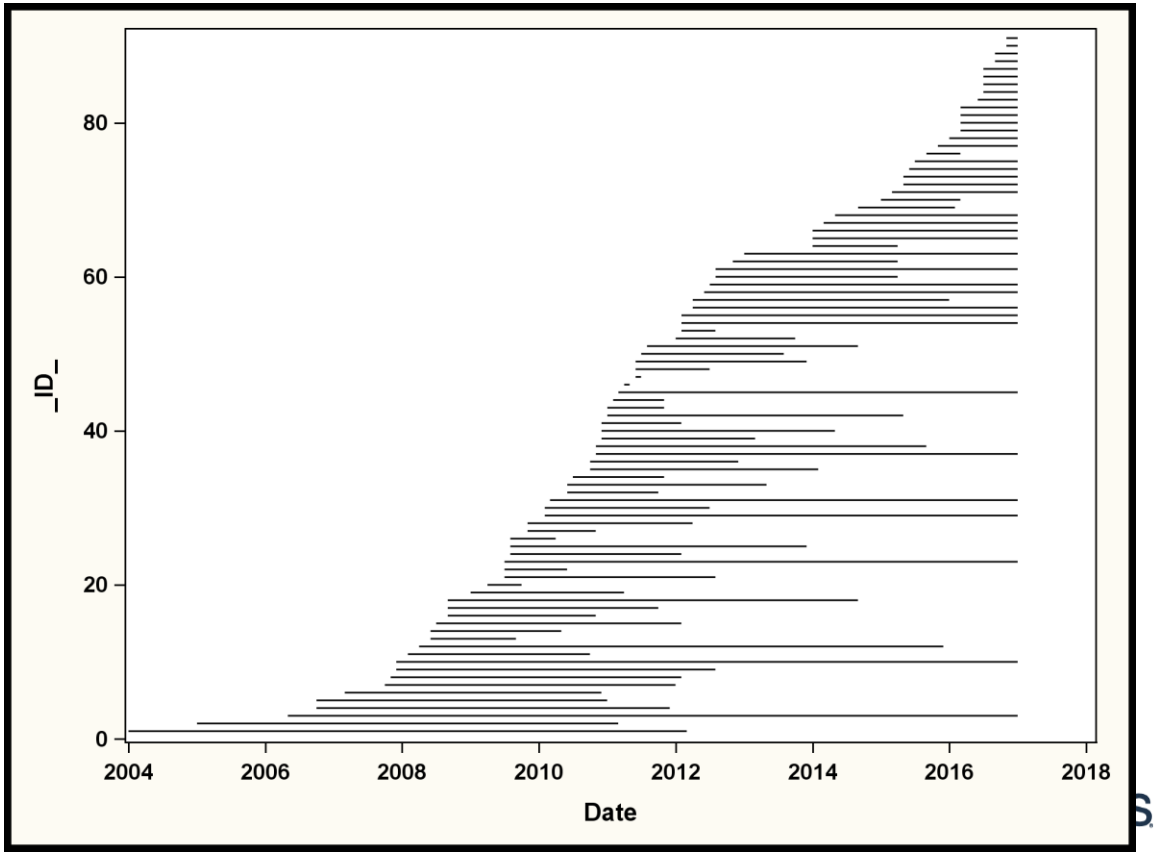
- Mitarbeiter werden Stichtag 01/2009 betrachtet
- Unternehmen wurde 01/2004 gegründet
- „Pre-Selektion“ der Daten





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

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



„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

Analyse von Input-Variablen mit der PHREG Procedure

Class Level Information					
Class	Value	Design Variables			
Department	ADMINSTRATION	-1	-1	-1	-1
	MARKETING	1	0	0	0
	SALES_ENGINEER	0	1	0	0
	SALES_REP	0	0	1	0
	TECH_SUPPORT	0	0	0	1
Gender	F	-1			
	M	1			
TechKnowHow	NO	-1			
	YES	1			

Analysis of Maximum Likelihood Estimates							
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Department	MARKETING	1	-1.15513	0.47794	5.8414	0.0157	0.606
Department	SALES_ENGINEER	1	0.82336	0.52244	2.4838	0.1150	4.380
Department	SALES_REP	1	0.62976	0.29224	4.6436	0.0312	3.609
Department	TECH_SUPPORT	1	0.35572	0.29940	1.4117	0.2348	2.744
TechKnowHow	YES	1	-0.63474	0.27370	5.3781	0.0204	0.281

Variable (Category)	Coefficient	p-Value
Department MARKETING	-1.155	0.016
Department SALES_ENGINEER	0.823	0.115
Department SALES_REP	0.630	0.031
Department TECH_SUPPORT	0.356	0.235
Department ADMIN	-0.654	
TechKnowHow YES	-0.635	0.020

$$-[(-1.155)+0.823+0.630+0.356] = -0.654$$

```
PROC PHREG DATA=Employees;
 CLASS department gender TechKnowHow / PARAM=effect REF=first;
 MODEL Duration*Status(1)= department gender TechKnowHow /
                           SELECTION=stepwise;
RUN;
```

Analyse der „Explained Variation“ mit der PHREG Procedure

Verwende die „EV“ Option im PHREG Statement

```
PROC PHREG DATA=Employees EV;
 CLASS department gender TechKnowHow/ PARAM=effect REF=first;
 MODEL Duration*Status(1)= department gender TechKnowHow;
RUN;
```

Predictive Inaccuracy and Explained Variation		
Predictive Inaccuracy (Smaller is Better)		Percent
Without Covariates	With Covariates	Explained Variation
0.3600	0.2921	18.84

Variables in the Model	Explained Variation
Department	13.7 %
TechKnowKow	2.0 %
Department, TechKnowKow	17.2 %
Department, TechKnowKow, Gender	18.4 %

Vorhersage der Survival mit der PHREG Procedure

```
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;
```

	EmpNo	Department	Gender	Start	End	Status	TechKnowHow	Duration	ObsAtRsik	SurvPred	DisplacmLikelihood
1	1001	MARKETING	M	01JAN2004	01MAR2012	0	NO	98	3	0.3000662358	0.0342095828
2	1002	SALES_REP	M	01JAN2005	01MAR2011	0	NO	74	9	0.0152709689	0.3883756359
3	1003	TECH_SUPPORT	M	01MAY2006	.	1	YES	128	1	0.2160216763	0.1379484728
4	1004	TECH_SUPPORT	M	01OCT2006	01DEC2011	0	YES	62	12	0.4732932188	0.0030581462
5	1005	SALES_ENGINEER	M	01OCT2006	01JAN2011	0	YES	51	25	0.4301588168	0.0038343292
6	1006	ADMINSTRATION	F	01MAR2007	01DEC2010	0	NO	45	28	0.5577228186	0.0137046542
7	1007	ADMINSTRATION	F	01OCT2007	01JAN2012	0	NO	51	25	0.5038628656	0.0073694734
8	1008	SALES_REP	M	01NOV2007	01FEB2012	0	NO	51	25	0.0842551576	0.0501603566
9	1009	ADMINSTRATION	F	01DEC2007	01AUG2012	0	NO	56	17	0.4368117997	0.007699317
10	1010	TECH_SUPPORT	M	01DEC2007	.	1	YES	109	2	0.2160216763	0.1379484728
11	1011	TECH_SUPPORT	M	01FEB2008	01OCT2010	0	NO	32	41	0.3835067447	0.0013479246
12	1012	MARKETING	M	01APR2008	01DEC2015	0	NO	92	4	0.3978245623	0.0317086643
13	1013	SALES_REP	M	01JUN2008	01SEP2009	0	NO	15	63	0.6635274122	0.0094219574
14	1014	SALES_REP	M	01JUN2008	01MAY2010	0	NO	23	52	0.5451210355	0.0056891192
15	1015	TECH_SUPPORT	M	01JUL2008	01FEB2012	0	YES	43	29	0.6640788277	0.0379949482

SAS Viya PHSELECT Procedure:
The CODE statement generates SAS code that predicts the survival function at specified years

```
proc phselect data=mycas.Customers;
   class Area(ref='Urban') LifeChange(ref='None')
         PlanType(ref='B') Satisfaction(ref='Poor')
          Smoking(ref='No') / param=ref;
   model Time*Status(0) = Age Area Education Income
                          LifeChange PlanType
                          Satisfaction Smoking;

   selection method=forward(select=bic stop=bic);
   code file='ScoreCode.txt' timepoint=12 24 36 48 60;
run;
```

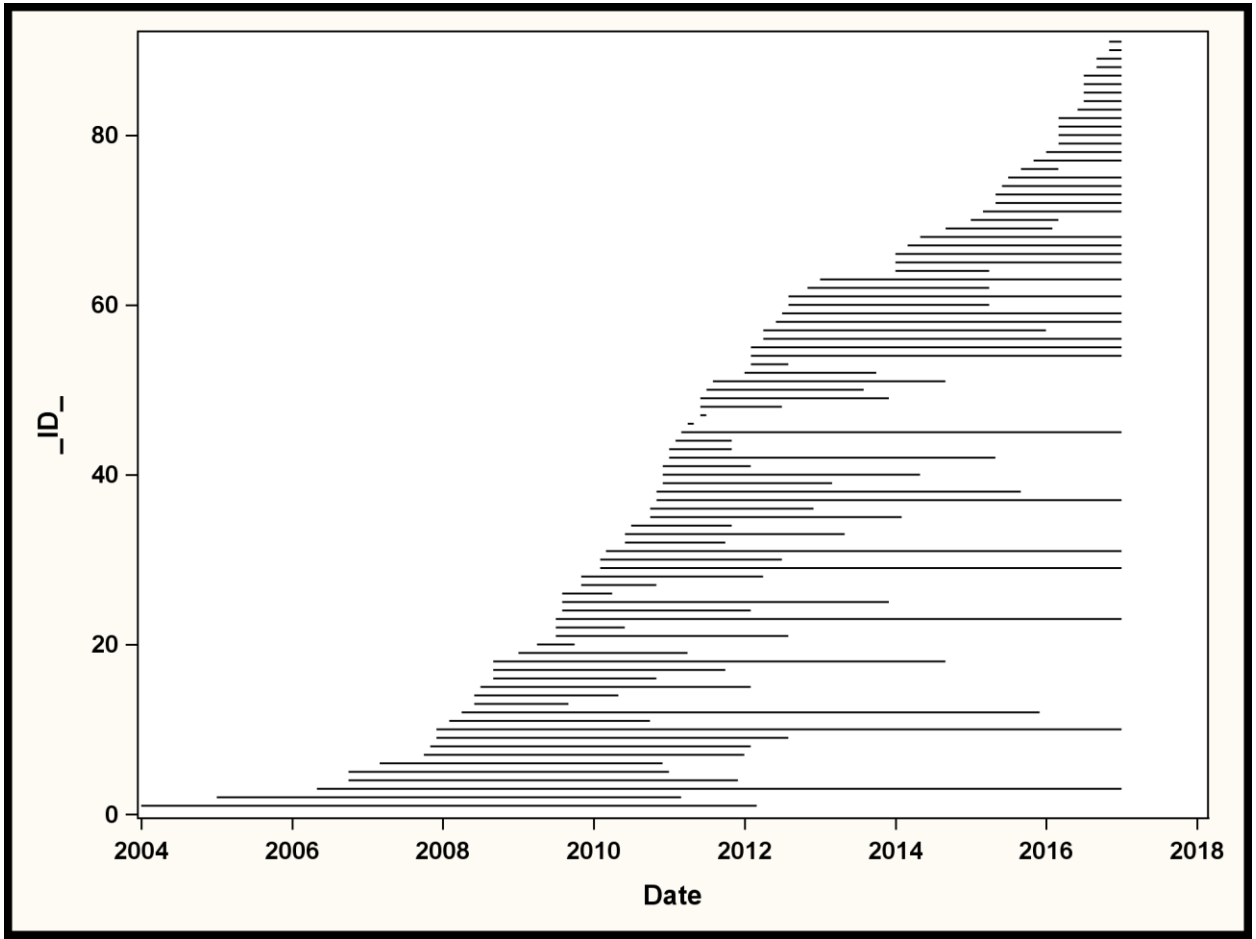
The score code predicts retention probabilities for new customers at the specified years

```
data Retention;
   set NewCustomers;
   %include 'ScoreCode.txt';
run;
```

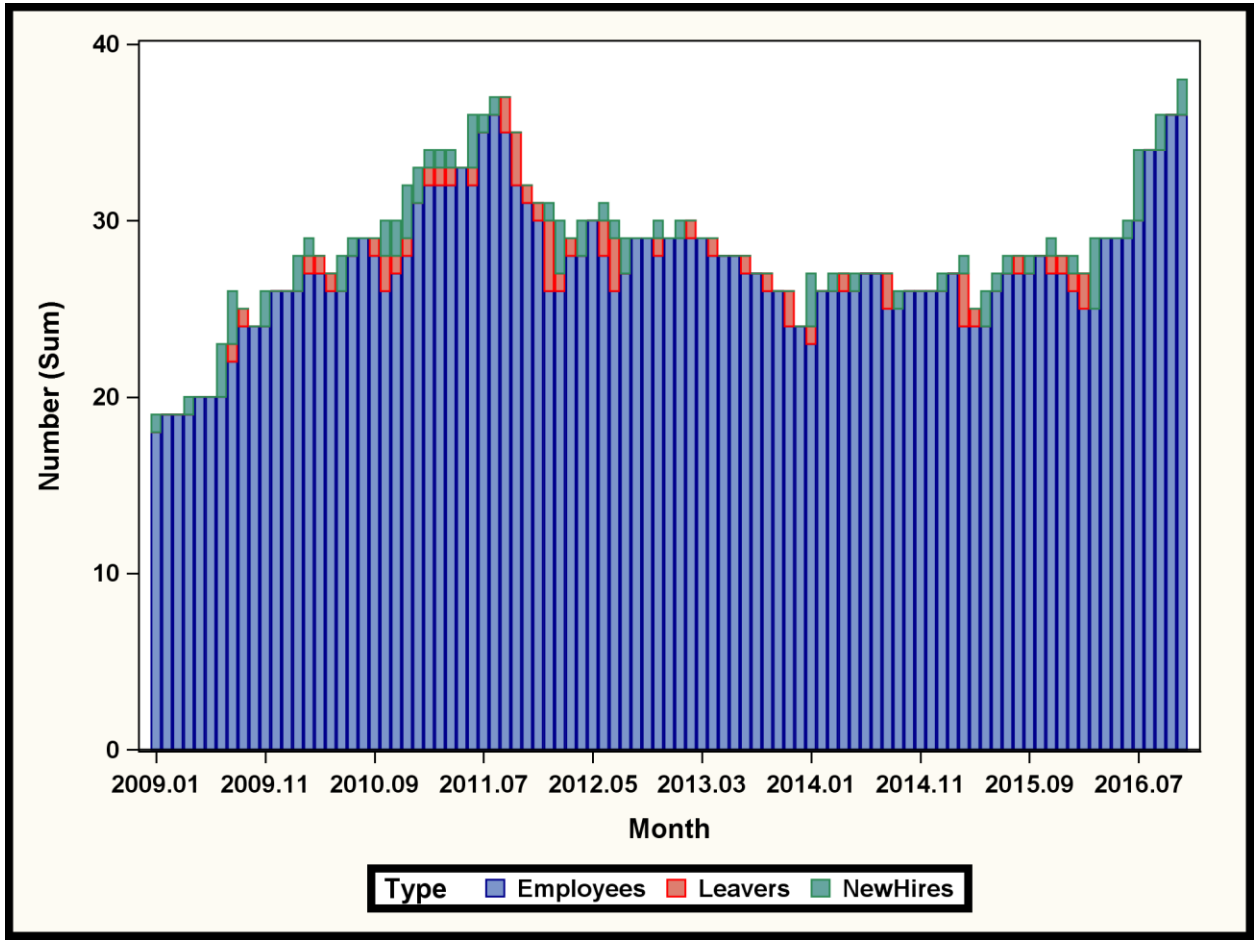
Area	Satisfaction	Life Change	Years of Education	Retention Probability at 1 Year	Retention Probability at 2 Years	Retention Probability at 3 Years	Retention Probability at 4 Years	Retention Probability at 5 Years
Rural	Poor	New Job	13	0.671	0.455	0.315	0.221	0.155
Urban	Good	Married	14	0.718	0.520	0.383	0.285	0.212
Rural	Excellent	New Job	8	0.711	0.512	0.373	0.276	0.204
Urban	Poor	New Job	11	0.652	0.431	0.290	0.198	0.136
Rural	Excellent	Child	17	0.786	0.622	0.498	0.402	0.324

Ausgewählte Plots für die Employee Survival

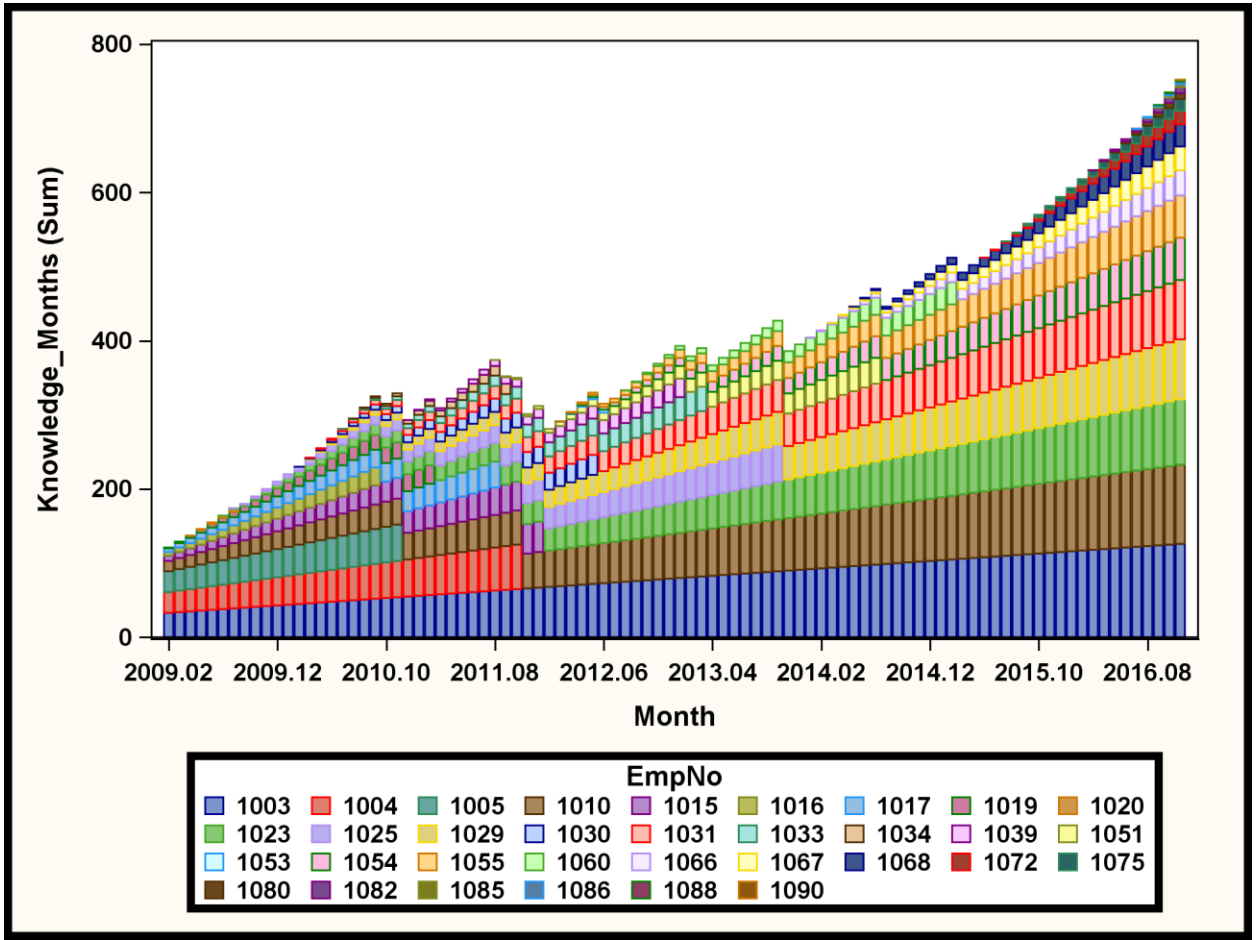
Career Start-End Plot



Employees-Win-Loss-Plot



Cumulated-Knowledge Plot



Zusammenfassung

- Die „Survival-Analyse“ ist auch in anderen Bereichen als der Medizin-Statistik oder der Analyse klinischer Studien sehr gut einsetzbar
- Survival Kurven können visualisiert und gut interpretiert werden

- SAS STAT Procedures und der SAS Enterprise Miner bieten Möglichkeiten zur Analyse von Ereignisdaten (zensierte Daten)

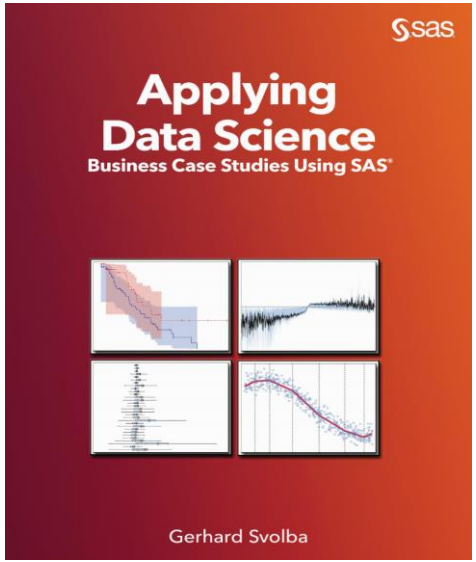
- Die Cox-Proportional Hazards Regression erlaubt die Identifikation und Bewertung von Einflussvariablen.

Links

- https://support.sas.com/en/books/authors/gerhard-svolba.html
- https://www.sas.com/store/books/categories/usage-and-reference/applying-data-science-business-case-studies-using-sas-/prodBK_63165_en.html
- Programme und Datasets: derzeit noch auf sascommunity.org werden demnächst nach github.com migriert.
- *AS/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. – Annals of Internal Medicine, 2001 Volume 136-10
- Redelmaier et al: Survival in Academy Award-Winning Actors and Actresses
- Sylvestre et. Al: Do Osca Winners Live Longer than Less Successful Peers? A Re-analysis of the Evidence – Annals of Internal Medicine, 2006, Volume 145-5

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