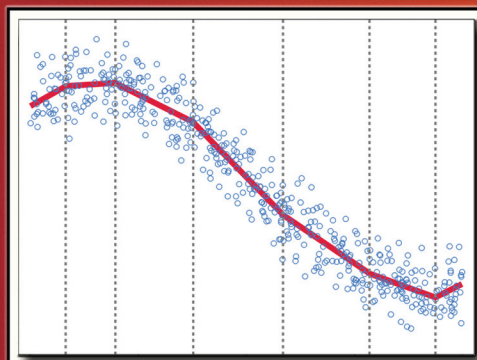
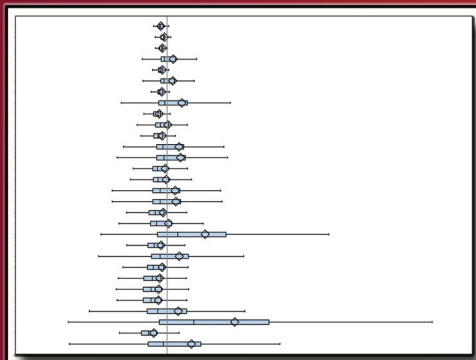
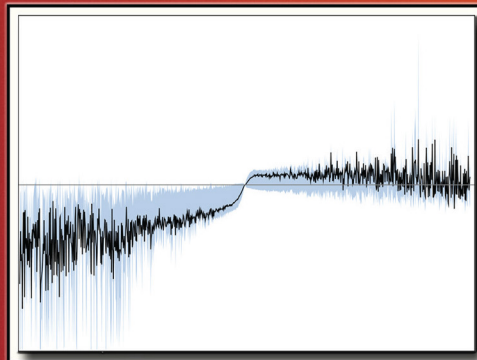
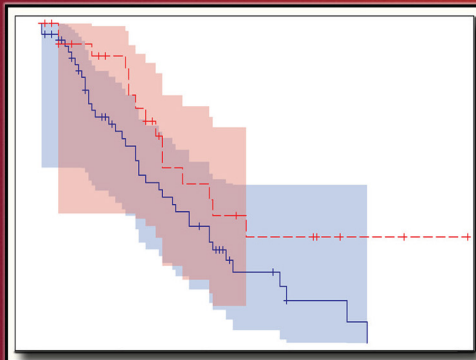
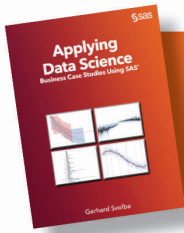


# Applying Data Science

## Business Case Studies Using SAS®



Gerhard Svolba



From *Applying Data Science*.  
Full book available for purchase [here](#).

# Contents

<b>Dedication .....</b>	<b>iii</b>
<b>About This Book .....</b>	<b>ix</b>
<b>About The Author .....</b>	<b>xxvix</b>
<b>Acknowledgments .....</b>	<b>xxxix</b>
<b>Case Study 1 – Performing Headcount Survival Analysis for Employee Retention .....</b>	<b>1</b>
<b>Chapter 1: Using Survival Analysis Methods to Analyze Employee Retention Time .....</b>	<b>3</b>
1.1 Introduction .....	3
1.1.1 Time-to-Event Data .....	3
1.1.2 Analytical Methods for Time-to-Event Data .....	4
1.2 Overview of the Case Study.....	4
1.3 Business Background and Business Question.....	4
1.3.1 Business Background.....	4
1.3.2 Business Questions.....	5
1.3.3 Employee Retention Data.....	5
1.4 Simple Descriptive Statistics Do Not help .....	7
1.5 The Kaplan-Meier Method Can Deal with Censored Data .....	8
1.5.1 The Basic Idea .....	8
1.5.2 Analyzing the Individual Duration .....	9
1.5.3 Code Example.....	10
1.5.4 Graphical Representation of the Kaplan-Meier Curve .....	11
1.6 Detailed Analysis of the Survival Curve.....	12
1.6.1 Creating the Survival Curve for All Employees .....	12
1.6.2 Interpreting the Survival Curve .....	14
1.6.3 Adding Confidence Bands to the Survival Curve .....	15
1.7 Interpreting the Hazard Curve.....	16
1.7.1 Basic Idea of the Hazard Curve .....	16
1.7.2 Adding a Plot for the Hazard Curve.....	16
1.7.3 The Hazard Curve for the SALES_ENGINEER Department .....	17
1.8 Additional Methods in PROC LIFETEST .....	18
1.8.1 Using the Lifetable Method .....	18
1.8.2 Generating an Output Data Set.....	20
1.9 Conclusion.....	21
<b>Chapter 2: Analyzing the Effect of Influential Factors on Employee Retention Time .....</b>	<b>23</b>

2.1 Introduction .....	23
2.2 Analyzing the Employee Data by Department .....	23
2.2.1 Descriptive Results .....	23
2.2.2 Survival Analysis.....	25
2.3 Additional Stratified Analyses.....	31
2.3.1 Survival Analysis by Gender and Technical Knowledge .....	31
2.3.2 The Misleading Effect of Left Truncated Data .....	33
2.4 Quantifying the Effect of Influential Variables .....	34
2.4.1 The Cox Proportional Hazards Regression .....	34
2.4.2 Results of the Cox Proportional Hazards Regression.....	36
2.4.3 Explained Variation of the Cox Proportional Hazards Model .....	37
2.4.4 Creating Output Data Sets .....	38
2.5 Preparing Time-to-Event Data.....	40
2.5.1 General Points .....	40
2.5.2 Business Decisions for the Definition of Events.....	40
2.6 Other Procedures in SAS/STAT® for the Analysis of Time-to-Event Data.....	41
2.7 Conclusion.....	42
<b>Chapter 3: Performing Survival Data Mining - The Data Mining Approach for Survival Analysis.....</b>	<b>43</b>
3.1 Introduction .....	43
3.2 The Idea of Survival Data Mining.....	44
3.2.1 Using a Multinomial Logistic Regression .....	44
3.2.2 Expanding the Data .....	44
3.3 Using SAS Enterprise Miner for Survival Data Mining .....	45
3.3.1 Using the Survival Node .....	45
3.3.2 Survival Data Mining Example - Preparation.....	45
3.3.3 Survival Data Mining Example – Results.....	48
3.4 Expanding Data for Survival Data Mining.....	52
3.4.1 Rationale .....	52
3.4.2 Different Data Requirements between SAS Enterprise Miner and SAS/STAT Procedures.....	52
3.4.3 Two Types of “Expanded” Data.....	52
3.5 Code Example to Expand Data for Survival Data Mining .....	54
3.5.1 General .....	54
3.5.2 Using a DATA Step Loop to Expand the EMPLOYEES Data.....	54
3.5.3 Merging a Multiple-Row-per-Subject Data Set to Expand the Employees Data .....	55
3.5.4 Be Careful with the Alignment of the Data! .....	57
3.6 Conclusion.....	58
<b>Chapter 4: Visualizing Employee Retention Data .....</b>	<b>59</b>
4.1 Introduction and Overview.....	59
4.2 Initial Preparations for the Reports.....	60
4.2.1 General Points .....	60
4.2.2 Initialization Statements .....	60

4.3 The Career-Start-End Plot .....	60
4.3.1 General Idea .....	60
4.3.2 The Graph and Its Interpretation .....	61
4.3.3 Data Preparation for the Career-Start-End Plot .....	61
4.4 Employees-Win-Loss-Plot .....	63
4.4.1 General Idea .....	63
4.4.2 The Graph and Its Interpretation .....	63
4.4.3 Preparing the data .....	64
4.5 Cumulated-Knowledge Plot .....	67
4.5.1 General Idea .....	67
4.5.2 The Graph and Its Interpretation .....	67
4.5.3 Preparing the Data .....	69
4.6 Conclusion .....	70
<b>Case Study 2 – Detecting Structural Changes and Outliers in Longitudinal Data ....</b>	<b>71</b>
<b>Chapter 5: Analyzing and Smoothing the Course of Longitudinal Data .....</b>	<b>73</b>
5.1 Introduction .....	73
5.2 Analyzing Longitudinal Data .....	74
5.2.2 Methods for Detection of Changes in Longitudinal Data .....	74
5.2.3 Automatic Detection of Changes and Outliers .....	75
5.3 Airline Passengers Data from 1990 and 2004 .....	75
5.3.1 The Data .....	75
5.3.2 Business Questions .....	76
5.4 Considerations When Smoothing Longitudinal Data .....	77
5.4.1 Information Reduction .....	77
5.4.2 Length of the Smoothing Window .....	77
5.4.3 Weighting of the Periods .....	77
5.4.4 Orientation of the Smoothing Window .....	77
5.5 Smoothing Your Data with the EXPAND Procedure .....	78
5.5.1 General .....	78
5.5.2 Backward Smoothing .....	78
5.5.3 Centered Smoothing .....	79
5.5.4 Specifying the Moving Weights Manually .....	81
5.5.5 Trimming in Case of Incomplete Data .....	81
5.6 Smoothing Longitudinal Data with a SAS DATA Step .....	84
5.6.1 Rationale for a SAS DATA Step Implementation .....	84
5.6.2 Brute-Force Method .....	84
5.6.3 Efficient Method .....	85
5.6.4 Calculating a Centered Moving Average with a DATA Step .....	86
5.6.5 Using a User-Defined SAS Function .....	86
5.7 Conclusion .....	87
<b>Chapter 6: Detecting Structural Changes in Longitudinal Data .....</b>	<b>89</b>
6.1 Introduction .....	89
6.2 Detect Structural Changes with Analytical Methods .....	90
6.2.1 Multivariate Adaptive Regression Splines .....	90

6.2.2 Automatic Detection of Knot Values .....	90
6.3 Using the ADAPTIVEREG Procedure .....	92
6.3.1 General .....	92
6.3.2 Basic Usage Examples .....	92
6.4 Breakpoints in the Airline Passenger Data .....	94
6.4.1 Interpreting the Breakpoints .....	94
6.4.2 Displaying the Breakpoints in the Line Chart.....	96
6.5 Structural Changes in Clinical Trial Data .....	98
6.5.1 Business Background and Data .....	98
6.5.2 Patient Recruitment over Time .....	99
6.5.3 Baseline Characteristics in a Clinical Trial .....	101
6.6 Conclusion.....	102
<b>Chapter 7 – Detecting Outliers and Level Shifts in Longitudinal Data .....</b>	<b>105</b>
7.1 Introduction .....	105
7.2 Detecting Outliers and Level Shifts .....	106
7.2.1 Overview.....	106
7.2.2 Level Shifts.....	108
7.2.3 Outliers .....	108
7.2.4 Other Event Types.....	109
7.3 Using SAS for the Detection of Events in Longitudinal Data .....	109
7.3.1 Procedures in SAS/ETS® Software.....	109
7.3.2 Methods in SAS® Forecast Server .....	110
7.4 Event Detection in the Airline Passenger Data.....	111
7.4.1 Overview.....	111
7.4.2 Detecting Events with the X13 Procedure.....	111
7.4.3 Creating a Plot with Reference Lines.....	112
7.5 Other Methods for Event Detection .....	115
7.5.1 Overview.....	115
7.5.2 Detecting Events with the X11 Procedure.....	115
7.5.2 Detecting Events with the HPFDIAGNOSE Procedure.....	116
7.6 Conclusion.....	117
<b>Chapter 8 – Results from a Simulation Study with Longitudinal Data .....</b>	<b>119</b>
8.1 Introduction .....	119
8.2 Simulating Longitudinal Data .....	119
8.2.1 Properties of the Simulated Data .....	119
8.2.2 Preparations for Simulating the Data.....	120
8.2.3 Using a DATA Step to Generate the Data.....	121
8.3 Detecting Structural Changes in the Simulated Data .....	123
8.3.1 Analyzing the Original Data .....	123
8.3.2 Detecting Changes in the Smoothed Data .....	124
8.4 Detecting Events in the Simulated Data.....	126
8.4.1 Using the X12 Procedure.....	126
8.4.2 Interpretation .....	127

8.5 Conclusion.....	128
<b>Chapter 9 – Analyzing the Variability of Longitudinal Data.....</b>	<b>129</b>
9.1 Introduction .....	129
9.2 Analyzing Variability over Time.....	130
9.2.1 Visual Interpretation of the Data.....	130
9.2.2 The Effect of Too Much Variability .....	131
9.3 Monitoring Longitudinal Data with Control Charts .....	132
9.3.1 Method from Statistical Process Control .....	132
9.3.2 Detecting Structural Changes in the X- and S- Curves.....	133
9.4 Misleading Changes and Outliers .....	135
9.4.1 Alternating Differences between March and April.....	135
9.4.2 Decline in Sales Numbers between January and April .....	136
9.5 Conclusion.....	136
9.6 Conclusion of This Case Study .....	137
<b>Case Study 3 – Explaining Forecast Errors and Deviations .....</b>	<b>139</b>
<b>Chapter 10 – Investigating Forecast Errors with Descriptive Statistics .....</b>	<b>141</b>
10.1 Introduction .....	141
10.2 Business Questions and Background .....	142
10.2.1 The Business Environment of the Case Study .....	142
10.2.2 Measuring the Forecast Error .....	144
10.2.3 Business Questions for the Analysis.....	144
10.3 Data Preparation.....	145
10.3.1 Available Source Data.....	145
10.3.2 Merging the Tables Together.....	147
10.3.3 Calculating Derived Variables.....	148
10.5 Descriptive Analysis of the Forecast Errors .....	150
10.5.1 Overview.....	150
10.5.2 Checking the Distribution of the Forecast Error.....	150
10.5.3 Using Box Plots to Analyze the Forecast Error by Subgroup.....	153
10.5.4 Analyzing Target Year and Model Type .....	158
10.5.5 Band Charts with Quartiles .....	161
10.6 Conclusion.....	163
<b>Chapter 11 – Investigating Forecast Errors with General Linear Models.....</b>	<b>165</b>
11.1 Introduction .....	165
11.2 Regression Analysis in SAS .....	165
11.2.1 Procedures in SAS/STAT® .....	165
11.2.2 General and Generalized Might Confuse You .....	166
11.2.3 Similar Procedures and Their Focus .....	166
11.3 Regression Analysis .....	167
11.3.1 General Points on Regression Models.....	167
11.3.2 Univariate Regression Analysis .....	169
11.4 Multivariate Regression Analysis .....	172
11.4.1 Using the GLMSELECT Procedure .....	172

11.4.2 Results from the Regression Model.....	173
11.4.3 Comparison with the Log-Transformed Model.....	174
11.4.4 Coefficients of the Variables in the Regression Model.....	174
11.5 Scoring New Observations with the GLMSELECT Procedure.....	176
11.6 Conclusion.....	178
<b>Chapter 12 – Interpreting the Coefficients of Categorical Variables in Regression Models .....</b>	<b>179</b>
12.1 Introduction.....	179
12.2 Categorical Variables and Dummy Variables.....	179
12.2.1 Problem Description.....	179
12.2.2 Different Types of Dummy Variable Coding.....	180
12.2.3 Defining Dummy Variables in Selected SAS Regression Procedures.....	181
12.3 Displaying Regression Coefficients for All Categories.....	183
12.3.1 Introduction.....	183
12.3.2 Usage Example.....	184
12.3.3 Macro Code Explained Step-by-Step.....	185
12.3.4 Summary.....	189
12.4 Imputing Missing Values with the GLM Procedure.....	189
12.4.1 Missing Values.....	189
12.4.2 A Code Example for Imputing Missing Values.....	190
12.5 Further Topics for SAS Regression Procedure.....	192
12.5.1 Saving Tables with the ODS OUTPUT Statement.....	192
12.5.2 Changing the Selection Statistic for the Stepwise Selection.....	193
12.5.3 Creating a Multi-Step Model.....	193
12.5 Conclusion.....	194
<b>Chapter 13: Using Quantile Regression to Get More Than the Average Picture ...</b>	<b>195</b>
13.1 Introduction.....	195
13.2 Basic Idea of Quantile Regression.....	196
13.2.1 Introduction.....	196
13.2.2 SAS Procedures for Quantile Regression.....	197
13.2.3 Robustness of the Quantile Regression.....	198
13.2.4 Model for the 0.25 Quantile Compared.....	200
13.3 Quantile Regression for the Statistical Forecast Error.....	202
13.3.1 Introduction.....	202
13.3.2 Quantile Regression for Selected Quartiles.....	202
13.3.3 Using the HPQUANTSELECT Procedure.....	204
13.3.4 Creating a Process Plot for the Parameter Estimates.....	205
13.4 Sampling of Data.....	206
13.4.1 Overview.....	206
13.4.2 Sampling with the SAS DATA Step.....	207
13.4.3 Sampling with the SURVEYSELECT Procedure.....	207
13.5 Conclusion.....	208
<b>Chapter 14 – Analyzing the Effect of Manual Overrides in Forecasting.....</b>	<b>209</b>

14.1 Introduction .....	209
14.2 Manual Overrides.....	209
14.2.1 Available Data.....	209
14.2.2 Quantifying the Effect of Manual Overrides .....	210
14.3 Univariate Results.....	212
14.3.1 Differences between Judgmental and Statistical Forecast.....	212
14.3.2 Analyzing Selected Variables.....	216
14.4 Results from the Regression Model .....	219
14.4.1 Introduction.....	219
14.4.2 Investigating Only the Size and Direction of the Manual Overwrite .....	219
14.4.3 Adding More Explanatory Factors.....	221
14.5 Conclusion.....	224
<b>Case Study 4 – Forecasting the Demand for New Products .....</b>	<b>225</b>
<b>Chapter 15 – Performing Demand Forecasting for New Products .....</b>	<b>227</b>
15.1 Introduction to the Case Study .....	227
15.2 Introduction .....	228
15.3 Business Questions and Background .....	228
15.3.1 Definition of Short Demand History .....	228
15.3.2 Reasons for Short or No Demand History .....	229
15.3.3 New Product or Short-Term Product.....	229
15.4 Analytic Approaches for Forecasting New Products.....	230
15.4.1 Overview.....	230
15.4.2 Demand Forecasting Using Predictive Modeling .....	230
15.4.3 Using Similarity Search for New Product Forecasting.....	230
15.4.4 Deriving Forecasts from Similarity Clusters.....	230
15.4.5 Using Time Series Similarity Analysis .....	230
15.5 Data Preparation Steps .....	231
15.5.1 Available Data .....	231
15.5.2 Generated Output Data.....	232
15.5.3 Overview of the Data Preparation Steps .....	232
15.5.4 Number of Products Launched in the Same Month .....	233
15.5.5 Aggregating the Demand Data .....	235
15.6 Creating the Demand Data Mart .....	236
15.6.1 Overview.....	236
15.6.2 Basic Variables .....	236
15.6.3 Adding the Known Demand .....	237
15.7 Adding Known Demand Data to Product Base Data .....	239
15.7.1 Introduction.....	239
15.7.2 Demand Data in One Row per Product.....	239
15.8 Conclusion.....	240
<b>Chapter 16 – Using Poisson Regression to Forecast the Demand for New Products.....</b>	<b>241</b>
16.1 Introduction.....	241



16.2 Poisson Regression.....	241
16.2.1 Generalized Linear Model.....	241
16.2.2 Implementation in SAS .....	242
Overview.....	242
16.3 Demand Forecasting Using Predictive Modeling .....	242
16.3.1 Basic Idea.....	242
16.3.2 Analysis Data .....	242
16.3.3 Independency of Observations.....	243
16.4 Poisson Regression Using the HPGENSELECT Procedure .....	244
16.4.1 SAS Code .....	244
16.4.2 Results of the Poisson Regression .....	245
16.4.3 Extending the Model Syntax .....	247
16.5 Scoring a New Product .....	248
16.5.1 Data for the New Product.....	248
16.5.2 Applying the Score Logic .....	249
16.5.3 Forecast Report.....	249
16.5.4 Graphical Output .....	250
16.6 Conclusion.....	251
<b>Chapter 17 – Using Similarity Search to Forecast the Demand for New Products.....</b>	<b>253</b>
17.1 Introduction .....	253
17.2 Basic Idea of Similarity Search.....	253
17.2.1 Basic Idea.....	253
17.2.2 Advantage of Similarity Search.....	254
17.3 Implementing Similarity Search .....	254
17.3.1 Product Repository .....	254
17.3.2 New Product Data .....	254
17.3.3 Defining the Weights.....	255
17.3.4 Similarity Search with the SQL Procedure .....	255
17.3.5 Preparation of the Results.....	257
17.3.6 Displaying the Demand Data.....	258
17.4 Extending the Similarity Search .....	260
17.4.1 Creating a SAS Stored Process .....	260
17.4.2 Similarity Search for Existing Products .....	261
17.5 Conclusion.....	261
<b>Case Study 5 – Checking the Alignment with Predefined Pattern.....</b>	<b>263</b>
<b>Chapter 18 – Checking Accounting Data for the Benford’s Law.....</b>	<b>265</b>
18.1 Introduction to the Case Study .....	265
18.1.1 General Idea and Examples.....	265
18.1.2 Statistical Implementation.....	266
18.2 Business Questions and Background .....	266

18.2.1 The Business Environment of the Case Study .....	266
18.2.2 The Available Data and Their Background .....	267
18.2.3 Business Questions for the Analysis of Benford's Law .....	268
18.3 Benford's Law .....	269
18.3.1 General Idea.....	269
18.3.2 Calculating the Expected Benford Proportions .....	269
18.3.3 Discussion of the Benford's Law.....	270
18.4 Checking Data for the Benford's Law .....	271
18.4.1 Using the Chi <sup>2</sup> -Distribution to Check Against the Expected Distribution .....	271
18.4.2 Deriving the First Digit from the Amount Value .....	272
18.4.3 Using the FREQ Procedure .....	272
18.5 Results .....	274
18.5.1 Preparation.....	274
18.5.2 Graphical Output .....	275
18.5.3 Tabular Output.....	275
18.5.4 Interpretation .....	276
18.6 Consideration of Missing Digits .....	276
18.6.1 Problem Description .....	276
18.6.2 Accounting for Zero Frequencies.....	277
18.6.3 Example for Data with No Digit 8.....	278
18.7 Conclusion.....	279
<b>Chapter 19 – Checking the Benford's Law for Multiple Accounts .....</b>	<b>281</b>
19.1 Introduction .....	281
19.2 Running Benford Analysis in BY-Entity Mode .....	281
19.2.1 Basic Idea.....	281
19.2.2 Implementation with the FREQ Procedure.....	282
19.2.3 Preparing a Table with Chi <sup>2</sup> -Values .....	283
19.3 Graphical Display of the Results .....	284
19.3.1 Overall Distribution of p-Values.....	284
19.3.2 Displaying Individual Courses.....	285
19.4 Accounting for Zero Frequencies .....	287
19.4.1 Problem Description .....	287
19.4.2 Using the TIMESERIES Procedure .....	287
19.5 Conclusion.....	289
<b>Chapter 20 – Checking Different Patterns in the Data .....</b>	<b>291</b>
20.1 Introduction .....	291
20.2 General Idea .....	291
20.2.1 Overview.....	291
20.2.2 Bank Account Data .....	292
20.3 Analysis by Time Period .....	292
20.3.1 Overall Results.....	292
20.3.2 Analysis per Year.....	293
20.3.3 Analysis per Month.....	294
20.3.4 Interpretation .....	296

20.4 Matching a Predefined Distribution .....	297
20.4.1 General Idea .....	297
20.4.2 Measures for Sales Targets .....	298
20.4.3 Ranking the Sales Agents .....	299
20.4.4 Investigating Individual Account Managers .....	301
20.5 Course of Customer Usage Data over Time .....	303
20.5.1 General Idea .....	303
20.5.2 Example Output .....	303
20.5.3 Summary .....	304
20.6 Conclusion .....	304

**Case Study 6 – Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods ..... 305**

**Chapter 21 – Finding Relationships in Your Analysis Data with Association Analysis ..... 307**

21.1 General Idea .....	307
21.1.3 Methods in SAS for Unsupervised Data Analysis .....	308
21.2 Business Question and Example Data .....	309
21.2.1 Association Analysis .....	309
21.2.2 Insurance Claim Data.....	309
21.3 Creating a Feature Data Mart from One-Row-Per-Subject Analysis Data .....	311
21.3.1 Overview .....	311
21.3.2 Using a SAS DATA Step.....	311
21.3.3 Alternate Approach: Using the TRANSPOSE Procedure .....	312
21.4 Performing Association Analysis with SAS® Enterprise Miner .....	314
21.4.1 Creating the SAS® Enterprise Miner Process Flow .....	314
21.4.2 Viewing the Results.....	316
21.4.3 Adding the p-Value for the Associations .....	318
21.5 Interpreting the Results.....	319
21.5.1 General .....	319
21.5.2 Interpreting the Rules .....	320
21.5.3 Obvious Results.....	320
21.5.4 Distant Associations .....	321
21.5.5 Analyzing Specific Subgroups .....	321
21.5.6 Visual Analysis of the Data .....	322
21.6 Conclusion.....	323

**Chapter 22 – Using Variable Clustering to Detect Relationships in Your Data ..... 325**

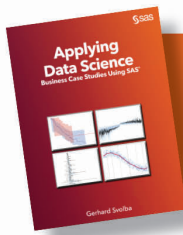
22.1 Introduction .....	325
22.2 Variable Clustering .....	326
22.2.1 Idea of Variable Clustering .....	326
22.2.2 Example Data.....	326
22.3 Preparing the Data.....	327
22.3.1 Create a Data Set with Indicator Variables .....	327

22.3.2 Method 1: Using the TRANSPOSE Procedure.....	327
22.3.3 Method 2: Using a SAS DATA Step.....	329
22.3.4 Method 3: Using the LOGISTIC Procedure.....	329
22.3.5 Conclusion .....	330
22.4 Performing Variable Clustering with the VARCLUS Procedure.....	330
22.4.1 General .....	330
22.4.2 Using the VARCLUS Procedure .....	331
22.4.3 Results of the VARCLUS Procedure .....	331
22.4.4 Creating a Tree Diagram .....	332
22.5 Interpretation of the Results.....	334
22.5.1 Reading the Tree Diagram .....	334
22.5.2 Creating and Reading the Variables Table .....	335
22.5.3 Interactive Analysis .....	335
22.6 Conclusion.....	336
<b>Chapter 23 – Investigating of Clinical Trial Data in an Explanatory Way .....</b>	<b>337</b>
23.1 Introduction .....	337
23.2 Patient Data from a Long-Term Clinical Trial .....	337
23.2.1 Basic Idea.....	337
23.2.2 Feature Categories in the Data.....	338
23.3 Preparing the Data for Association Analysis .....	339
23.3.1 Overview.....	339
23.3.2 Creating the Categories .....	339
23.3.3 Creating a Table with Features per Patient.....	341
23.4 Results of the Association Analysis.....	341
23.4.1 Obvious Rules Because of the Definition of the Data .....	342
23.4.2 Intuitive Rules from a Medical Point of View.....	342
23.4.3 Analyzing Specific Subgroups .....	343
23.5 Performing Variable Clustering .....	345
23.5.1 Introduction.....	345
23.5.2 Preparing the Data for Variable Clustering .....	345
23.5.3 Performing Variable Clustering.....	346
23.5.4 Interpretation of the Results .....	347
23.6 Conclusion.....	348
<b>Case Study 7 – Using Monte Carlo Simulations to Understand the Outcome Distribution .....</b>	<b>349</b>
<b>Chapter 24 – Calculating the Outcomes of All Possible Scenarios.....</b>	<b>351</b>
24.1 Overview over the Case Study .....	351
24.1.1 Introduction.....	351
24.1.2 Sales Manager’s Pipeline .....	352
24.2 Example Data and Business Question .....	352
24.2.1 The List of Sales Projects.....	352
24.2.2 Expectations of Upper Management .....	353
24.3 Some Basic Calculations .....	354

24.3.1 Not to Be Applied! – Incorrect Quick Calculation .....	354
24.3.2 Calculating the Expected Value .....	354
24.4 Complete Calculation .....	355
24.4.1 The Basic Idea .....	355
24.4.2 Some Basic Considerations .....	355
24.4.3 Looking at More Projects .....	356
24.5 Performing the Complete Calculation with SAS/IML Software .....	357
24.5.1 Introduction.....	357
24.5.2 General Idea of Using SAS IML for the Complete Calculations .....	357
24.5.3 Full SAS/IML Code .....	358
24.5.4 Details for the SAS/IML Code .....	358
24.6 Results from the Complete Calculation .....	362
24.6.1 Introduction.....	362
24.6.2 Results.....	363
24.6.3 Interpretation of the Results .....	364
24.7 Using a SAS DATA Step for the Complete Calculations.....	365
24.7.1 General Idea.....	365
24.7.2 SAS Code .....	365
24.8 Conclusion.....	368
<b>Chapter 25 – Using Monte Carlo Methods to Simulate the Distribution of the Outcomes .....</b>	<b>369</b>
25.1 Introduction .....	369
25.2 Performing Monte Carlo Simulations .....	369
25.2.1 Alternative to the Complete Calculation.....	369
25.2.2 Basic Idea.....	370
25.3 Using SAS/IML Software to Perform the Monte Carlo Simulations .....	370
25.3.1 Input Data.....	370
25.3.2 Full SAS/IML Code .....	370
25.3.3 SAS/IML Code in Details.....	371
25.4 Results of the Monte Carlo Simulations.....	373
25.4.1 Creating the Analysis Reports .....	373
25.4.2 Results of Different Scenarios .....	374
25.5 Using a SAS DATA Step to Perform the Monte Carlo Simulations.....	376
25.5.1 General Idea.....	376
25.5.2 The SAS DATA Step Code .....	377
25.6 Conclusion.....	378
<b>Case Study 8 – Studying Complex Systems – Simulating the Monopoly Board Game .....</b>	<b>379</b>
<b>Chapter 26 – Creating a Basic Framework to Simulate the Visit Frequency on the Fields of the Monopoly Board Game .....</b>	<b>381</b>
26.1 Simulating Complex Systems.....	381
26.1.1 Understanding Complex Systems with Simulation Studies .....	381

26.1.2 Case Study: Monopoly Board Game .....	382
26.2 Overview of the Case Study.....	382
26.3 Description of the Monopoly Board Game.....	382
26.3.1 Introduction.....	382
26.3.2 The Idea of the Game.....	382
26.3.3 Assumptions for the Simulation Runs.....	384
26.4 Business Questions of Interest .....	385
26.5 Simulating the Monopoly Board Game with a SAS DATA Step .....	386
26.5.1 Overview.....	386
26.5.2 Simulation Architecture .....	386
26.6 The SAS DATA Step for the Simulation in Detail .....	387
26.6.1 Initialization .....	387
26.6.2 Looping over Scenarios (Games).....	389
26.6.3 Looping over Rounds and Players.....	389
26.6.4 Considering the Go-to-Jail Directive .....	390
26.6.5 Closing the Macro Loop .....	391
26.7 Results from the Simulation Runs .....	391
26.7.1 Results in the Data Set .....	391
26.7.2 Preparing the Analysis Data.....	392
26.7.3 Graphical Output of the Players' Positions .....	393
26.7.4 Tabular Output of the Players' Positions .....	394
26.7.5 Considering the "Go-to-Jail" Directive .....	395
26.8 Conclusion.....	396
<b>Chapter 27 – Enhancing the Simulation Framework to Consider Special Rules ....</b>	<b>397</b>
27.1 Overview .....	397
27.2 Considering the Chance and Community Chest Fields.....	397
27.2.1 Business Background.....	397
27.2.2 Considering Community Chest Fields .....	398
27.2.3 Considering Chance Fields .....	398
27.2.4 SAS Code to Generate the Results .....	399
27.2.5 Results and Interpretation.....	401
27.3 Considering the Speed Die .....	402
27.3.1 Instructions for the Speed Die .....	402
27.3.2 Effect of the Relocation Rule .....	403
27.3.3 Implementation of the Speed Die Rules .....	403
27.4 Results of Considering the Speed Die.....	405
27.4.1 Distribution of the Visit Frequency .....	405
27.4.2 Illustrating the Dynamic over the Course of the Game .....	407
27.5 Conclusion.....	411
<b>Chapter 28 – Simulating the Profitability of the Property Fields of the Monopoly Board Game.....</b>	<b>413</b>
28.1 Introduction .....	413
28.2 Relation to Business Decisions.....	413
28.3 Tracking Players' Positions and Transactions.....	414

<b>28.3 Implementation with a SAS DATA Step</b> .....	<b>417</b>
28.3.1 Using a SAS Format as Lookup Table.....	417
28.3.2 Calculating the Monetary Values in the Simulation .....	418
<b>28.4 Field Profit over the Course of the Game</b> .....	<b>420</b>
28.4.1 Preparing the Data .....	420
28.4.2 Results and Interpretation.....	421
<b>28.5 Analyzing the Final Field Profit and Field Setup</b> .....	<b>422</b>
28.5.1 Introduction.....	422
28.5.2 Field Profit .....	422
28.5.3 Field Setup .....	424
<b>28.6 Profit and Risk Analysis</b> .....	<b>425</b>
28.6.1 General Idea.....	425
28.6.2 Creation of the Results .....	425
28.6.3 Interpretation of the Results .....	426
<b>28.7 Conclusion</b> .....	<b>427</b>
<b>Appendix A – SAS Macros</b> .....	<b>429</b>
A.1 The %VAREXIST macro .....	429
A.2 The %DetectBrkPoint Macro .....	430
A.3 The %CALC_REFERENCE_CATEGORY Macro.....	432
A.4 The %REPLACE_MV Macro .....	434
A.5 The %Monopoly_Sim Macro .....	435
<b>References</b> .....	<b>441</b>
Recommended Reading.....	443
<b>Index</b> .....	<b>445</b>



From *Applying Data Science*.  
Full book available for purchase [here](#).

## Case Study 1 – Performing Headcount Survival Analysis for Employee Retention

### Example Business Question for This Case Study

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

### Analytical Methods and SAS Procedures Applied

Survival analysis methods like Kaplan-Meier estimates, Cox Proportional Hazards regression and Survival Data Mining are used to solve the business questions.

### Analytic SAS Procedures

LIFETEST

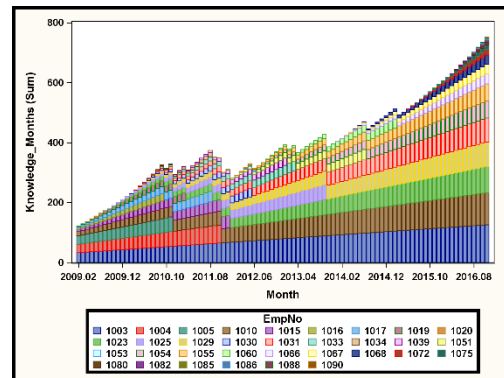
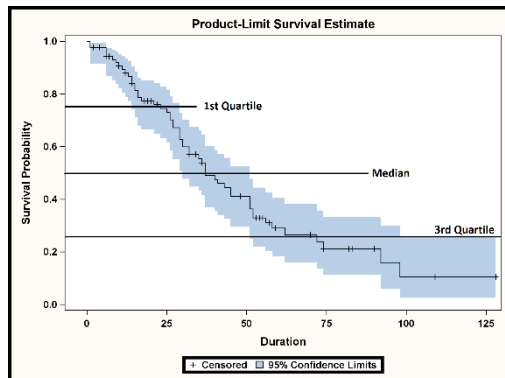
PHREG

Survival node in SAS® Enterprise Miner™

### Chapters in This Case Study

- Using Survival Analysis Methods to Analyze Employee Retention Time
- Analyzing the Effect of Influential Factors on Employee Retention Time
- Performing Survival Data Mining - The Data Mining Approach for Survival Analysis
- Visualizing Employee Retention Data

### Example Output





## **2** *Applying Data Science: Business Case Studies Using SAS*

# Chapter 1: Using Survival Analysis Methods to Analyze Employee Retention Time

<b>1.1 Introduction</b> .....	<b>3</b>
1.1.1 Time-to-Event Data .....	3
1.1.2 Analytical Methods for Time-to-Event Data .....	4
<b>1.2 Overview of the Case Study</b> .....	<b>4</b>
<b>1.3 Business Background and Business Question</b> .....	<b>4</b>
1.3.1 Business Background .....	4
1.3.2 Business Questions .....	5
1.3.3 Employee Retention Data .....	5
<b>1.4 Simple Descriptive Statistics Do Not help</b> .....	<b>7</b>
<b>1.5 The Kaplan-Meier Method Can Deal with Censored Data</b> .....	<b>8</b>
1.5.1 The Basic Idea.....	8
1.5.2 Analyzing the Individual Duration.....	9
1.5.3 Code Example .....	10
1.5.4 Graphical Representation of the Kaplan-Meier Curve .....	11
<b>1.6 Detailed Analysis of the Survival Curve</b> .....	<b>12</b>
1.6.1 Creating the Survival Curve for All Employees.....	12
1.6.2 Interpreting the Survival Curve.....	14
1.6.3 Adding Confidence Bands to the Survival Curve .....	15
<b>1.7 Interpreting the Hazard Curve</b> .....	<b>16</b>
1.7.1 Basic Idea of the Hazard Curve .....	16
1.7.2 Adding a Plot for the Hazard Curve.....	16
1.7.3 The Hazard Curve for the SALES_ENGINEER Department.....	17
<b>1.8 Additional Methods in PROC LIFETEST</b> .....	<b>18</b>
1.8.1 Using the Lifetable Method.....	18
1.8.2 Generating an Output Data Set .....	20
<b>1.9 Conclusion</b> .....	<b>21</b>

---

## 1.1 Introduction

---

### 1.1.1 Time-to-Event Data

The business question that is analyzed in this case study is taken from the human resources area. The retention time of employees is analyzed to generate results about the average length of the retention period and the effect of various influential factors.

The data for the case study are taken from a company that operates in the technical area. The company is the local operation of a larger brand and re-sells technical equipment for its mother company. Around 30 employees are responsible for the local market.

### Missing Endpoint and Censoring of Observations

This case study shows how analytical methods for survival analysis can be used to analyze time-to-event data. One specific feature of time-to-event data is that not all time intervals might be fully observed and the

## 4 *Applying Data Science: Business Case Studies Using SAS*

endpoint is unknown. In this case the mechanism of “censoring” of time intervals applies; intervals with no end date are cut at the last available date and this fact is specially treated in the analysis.

Consequently, two different types of time intervals enter the analysis:

- Intervals where the employee has left the company and the start and end time of his employment is known.
- Intervals for employees that are still with the company. Here the endpoint has not yet occurred and the only statement that can be made is that he has been with the company for a certain number of months.

---

### 1.1.2 Analytical Methods for Time-to-Event Data

In this case study it will be shown how the Kaplan-Meier method can be used to treat these different situations and to produce correct results. You will learn that conclusions about the average length of time intervals can be drawn, even if some of the endpoints have not yet been observed. You will also see that survival curves give you a clear visual impression of the distribution of the retention times of the employees.

Advanced analytical methods allow you to investigate the influence of different influential factors on the employment length, for example, by stratifying the analysis by different groups or by ranking these factors by their predictiveness for the employment duration.

Also, descriptive graphical methods can be a big help in learning from human resources data. The case study will also show advanced graphical methods to display the start and end of the various career or how the cumulative knowledge evolves over time.

---

## 1.2 Overview of the Case Study

The description of this case study extends over 4 chapters:

- This chapter explains the principles of the Kaplan-Meier method to analyze time-to-event data and illustrates this with survival curves and hazard curves on employee retention example data.
- Chapter 2 extends the concept of survival analysis to consider influential factors as stratification variables and as input variables for a regression model on survival data.
- Chapter 3 introduces how methods of survival data mining in SAS® Enterprise Miner™ can be used to analyze the employee retention data.
- Finally, specialized graphical methods for general analysis of employee data are shown in chapter 4.

---

## 1.3 Business Background and Business Question

### 1.3.1 Business Background

The data for this case study are taken from a company that operates in the technical area. The company is the local operation of a larger brand and re-sells technical equipment for its mother company. It is responsible for the local market and has currently 30 employees that work in the following departments of the company:

- **MARKETING:** advertising the company and its products on the market by running marketing campaigns on different channels and taking care about the public relations
- **SALES\_REP:** sales representatives that are responsible to sell the technical products to new and existing customers
- **SALES\_ENGINEER:** assisting the sales representatives in the sales process by doing sales presentations, product demonstrations, and covering the technical communication with prospective customers.

- **TECH\_SUPPORT**: technical experts that communicate with the customer in the post-sale phase by acting as a technical support hotline and assisting the customer with the introduction of the product in his company
- **ADMINISTRATION**: covering the back-office tasks of the company by providing functions like reception desk, accounting, legal, human resources, and office management.

---

### 1.3.2 Business Questions

Recently, an increasing number employees quit their job. Thus, the general manager of the company is interested to get a clearer picture about the average retention period of the employees and potential influential factors on the length of the retention period. The following questions are important to the manager from a business point of view:

- What is the average retention period for employees in the company?
- How can the retention period be visualized and compared between different subgroups?
- 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 groups?
- 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?
- What are the most relevant visualizations for this type of employee data?

Considering the fact that not all time intervals have an observed end date, the general manager understands that these analyses cannot just be made by comparing simple means of the length of the time intervals and is open to other methods.

---

### 1.3.3 Employee Retention Data

#### Base Data

The data that are presented in this chapter were recorded in the time interval January 2009 until December 2016. In this interval, 91 employees have been observed. For every employee the following variables have been recorded.

**Table 1.1: Variables in the EMPLOYEES Data Set**

Variable Name	Description
EmpNo	Employee Number
FirstName	First Name of the Employee
Gender	Employees' Gender
Department	Department, where the employee worked
TechKnowHow	Indicator, whether the employee has technical knowledge about the products
Start	Start date, when the employee started with the company
End	End date of the employees' employment. Missing, if he/she is still with the company

#### Censoring of the Retention Period

In Output 1.1 you see the data for employees 1021 – 1029. Consider the records of Frank (#1022) and Alan (#1023). Both started at July 2009. Frank left the company on June 2010, while Alan is still with the company when the analysis is performed on January 2017.

**Output 1.1: Selected Rows from the EMPLOYEES table**

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	Francesca	ADMINISTRATION	F	01AUG2009	01FEB2012	0	30
1025	Karl	SALES_ENGINEER	M	01AUG2009	01DEC2013	0	52
1026	Hana	ADMINISTRATION	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

Frank's time interval ends with an event (termination of employment). Alan's career did not end yet. We know only that he is still with the company when the analysis is performed. Consequently, Alan's observation periods need to be censored on January 2017.

This date is also called the censoring date. It denotes the point in time when the database has been closed and no information from later points in time is available.

- The derived variable STATUS has been created to indicate that the end date of a career is not observed, but the interval has been censored at a certain point in time, in this case on January 2017. In this case STATUS has the value 1; otherwise, it has the value 0.
- Variable DURATION describes the length of the time period for each employee. For those with an observed end date, DURATION is the interval length between start and end date. For those employees that are censored, DURATION describes the interval length between start date and censoring date.

Thus, the DURATION for Frank is 11 months indicating a known endpoint of the employment. Alan is still with the company. His DURATION is 90 months (7.5 years from July 2009 until January 2017) indicating the time when the last information about his employment is available.

The fact that the end date of the interval is unknown is also called "right censored". If the start value of the interval were missing, it would be called "left censoring".

### Left Truncation of Data

Data collection started in January 2009 and ended in December 2016. In 2009, however, the company has already existed for a couple of years. Thus, you can find employee records in the data for employees that were hired before 2009. As the data recording for the analysis only started on 2009, those employees that left the company before 2009 were not observed and are not recorded in the data.

**Output 1.2: First 19 Rows from the EMPLOYEES Table**

	EmpNo	FirstName	Department	Gender	Start	End	Status
1	1001	Don	MARKETING	M	01JAN2004	01MAR2012	0
2	1002	Hugh	SALES_REP	M	01JAN2005	01MAR2011	0
3	1003	Jim	TECH_SUPPORT	M	01MAY2006	.	1
4	1004	Art	TECH_SUPPORT	M	01OCT2006	01DEC2011	0
5	1005	Viktor	SALES_ENGINEER	M	01OCT2006	01JAN2011	0
6	1006	Petra	ADMINISTRATION	F	01MAR2007	01DEC2010	0
7	1007	Jana	ADMINISTRATION	F	01OCT2007	01JAN2012	0
8	1008	Peter	SALES_REP	M	01NOV2007	01FEB2012	0
9	1009	Susan	ADMINISTRATION	F	01DEC2007	01AUG2012	0
10	1010	Paul	TECH_SUPPORT	M	01DEC2007	.	1
11	1011	Carlos	TECH_SUPPORT	M	01FEB2008	01OCT2010	0
12	1012	Marius	MARKETING	M	01APR2008	01DEC2015	0
13	1013	Thomas	SALES_REP	M	01JUN2008	01SEP2009	0
14	1014	Bert	SALES_REP	M	01JUN2008	01MAY2010	0
15	1015	Robert	TECH_SUPPORT	M	01JUL2008	01FEB2012	0
16	1016	Dominique	TECH_SUPPORT	M	01SEP2008	01NOV2010	0
17	1017	Patricia	TECH_SUPPORT	F	01SEP2008	01OCT2011	0
18	1018	Karen	ADMINISTRATION	F	01SEP2008	01SEP2014	0
19	1019	Rainer	SALES_ENGINEER	M	01JAN2009	01APR2011	0

You see that the data represent a biased picture of the employee careers.

- Those who started before 2009 are documented in the data only if they stayed with the company at least until 2009.
- Those who left earlier are not in the sample.

This fact is called “left truncation”. Left truncation means that you get a biased picture for a period; only those employees who have an end date after a certain date are recorded in the data. The shorter periods (those who quit before) are not in the data. Chapter 2 shows methods to handle this situation.

For descriptive purposes and to define subgroups, a derived variable STARTPERIOD has been created. This variable groups the start date into the intervals: 2004–2008, 2009–2013, and 2014–2016. You see that the first group contains those hiring years from which only those employees are left, who are still active at the start of the data recording.

## 1.4 Simple Descriptive Statistics Do Not help

### Non-Observed Endpoints

Using simple descriptive statistics provides little help in getting insight into the average length of the retention period. Consider the records for the 11 employees in the “SALES\_ENGINEER” department shown in Output 1.3.

**Output 1.3: Department SALES ENGINEERS**

	EmpNo	FirstName	Department	Gender	Start	End	Status	Duration
1	1005	Viktor	SALES_ENGINEER	M	01OCT2006	01JAN2011	0	51
2	1019	Rainer	SALES_ENGINEER	M	01JAN2009	01APR2011	0	27
3	1020	John	SALES_ENGINEER	M	01APR2009	01OCT2009	0	6
4	1023	Alan	SALES_ENGINEER	M	01JUL2009	.	1	90
5	1025	Karl	SALES_ENGINEER	M	01AUG2009	01DEC2013	0	52
6	1030	Vincenz	SALES_ENGINEER	M	01FEB2010	01JUL2012	0	29
7	1055	Eugene	SALES_ENGINEER	M	01FEB2012	.	0	59
8	1060	George	SALES_ENGINEER	M	01AUG2012	01APR2015	0	32
9	1066	Mark	SALES_ENGINEER	M	01JAN2014	.	0	36
10	1082	Lucas	SALES_ENGINEER	M	01MAR2016	.	0	10
11	1086	Brady	SALES_ENGINEER	M	01JUL2016	.	0	6

- Six of them resigned and have an end date. These are the employees Viktor, Rainer, John, Karl, Vincenz, and George. Their duration has been simply calculated as the difference between start and end date.
- The other five employees, Alan, Eugene, Mark, Lucas, and Brady have no end date as they are still with the company. The retention periods have been censored and the duration has been calculated from the start date until January 2017. You see for example, that Brady has a duration of six months, which is the interval length between July 2016 and January 2017. The censoring status for these employees has been set to 1.

Output 1.4 shows the same data sorted by duration in ascending order.

**Output 1.4: Department SALES ENGINEERS Sorted by Duration**

EmpNo	FirstName	Department	Gender	Start	End	Status	Duration
1020	John	SALES_ENGIN...	M	01APR2009	30SEP2009	0	6
1086	Brady	SALES_ENGIN...	M	01JUL2016	.	1	6
1082	Lucas	SALES_ENGIN...	M	01MAR2016	.	1	10
1019	Rainer	SALES_ENGIN...	M	01JAN2009	31MAR2011	0	27
1030	Vincenz	SALES_ENGIN...	M	01FEB2010	30JUN2012	0	29
1060	George	SALES_ENGIN...	M	01AUG2012	31MAR2015	0	32
1066	Mark	SALES_ENGIN...	M	01JAN2014	.	1	36
1005	Viktor	SALES_ENGIN...	M	01OCT2006	31DEC2010	0	51
1025	Karl	SALES_ENGIN...	M	01AUG2009	30NOV2013	0	52
1055	Eugene	SALES_ENGIN...	M	01FEB2012	.	1	59
1023	Alan	SALES_ENGIN...	M	01JUL2009	.	1	90

### Need to Make Assumptions

In order to calculate an estimate for the average retention period, you could follow different approaches:

- Considering only records for employees that have an endpoint and for whom the variable END is not missing. This however means that you completely ignore the six observations that have been censored. In that case, the mean retention period is 32.8 months.
- Assuming that for the censored observations, the endpoint will immediately take place next month. This means you assume that the 5 employees that have not yet left, will resign right now. This is a very conservative assumption that has a mean retention period of 36.6 months.
  - For this calculation, the duration values of the non-observed endpoints (Status = 1) have been increased by 1 and the duration values of the observed endpoints have been used as they are.
  - Even if you make this “worst case” assumption, the average retention period is longer than the period from calculated in the first approach where obviously records with a long duration are ignored.
- You can create additional scenarios by making different assumption of the remaining retention period of those 5 employees who have been censored from the analysis.
  - Assuming on average 12 additional months until a termination of the employment, results in an average survival of 41.6 years. For this calculation the duration values of the non-observed endpoints (Status = 1) have been increased by 12 and the duration values of the observed endpoints have been used as they are.

You see that you won’t receive a satisfactory and interpretable solution with any of these assumptions and applying only basic descriptive statistics.

---

## 1.5 The Kaplan-Meier Method Can Deal with Censored Data

---

### 1.5.1 The Basic Idea

The Kaplan-Meier method can deal with the fact that not all employees’ careers have been observed until the endpoint. Over the range of individual retention times, the number of employees that are “at-risk” of leaving the company is calculated and used to weigh the number of events over time.

- At time 0, all employees are at risk of leaving the company.
- If the number of employees decreases over the duration time axis, the at-risk number is updated.

This allows you to calculate a weighted survival that can be interpreted as the proportion of employees surviving until a certain point in time.

## 1.5.2 Analyzing the Individual Duration

Table 1.2 shows the careers of the employees in the SALES\_ENGINEER department ordered by the duration of each individual career. The table is similar to the one shown in Output 1.3; it has however additional variables.

- Variable LEFT describes the number of employees that are still with the company at the end of the interval.
- Variables RESIGNED and CENSORED indicate how the respective records have been considered in the calculation for the survival estimate.
- Variable SURVIVAL holds the product limit survival estimate. You see that it only changes its value when the RESIGN variable equals 1. Compare this to Allison [1] for more details about the calculation of the survival estimates.

The DURATION column represents the amount of time with the company, up to the analysis date (January 2017). For example, the sales engineer with the most tenure has been with the company 90 months, and is still employed in January 2017 (thus his record is censored at event 90).

**Table 1.2: Results of the Kaplan-Meier Analysis**

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

Observe the following points in the table:

- The first line (duration 0) represents the start of the observation period. 11 employees are in the analysis.
- The next event takes place after a duration of 6 months, when John resigns. He was with the company from April 2009 until October 2009. Also, after 6 months, the observation of Brady has to be censored. He started his employment in July 2016. When the analysis takes place in January 2017, he has been 6 months with the company.
- At the beginning of the 6<sup>th</sup> month, 11 employees were observed. At the end of the 6<sup>th</sup> month there were 9 employees left (one event, one censored observation). One event took place and the Survival was computed accordingly.
- In month 10, no events take place but the observation of Lucas is censored. He started at March 2016.
- In month 27, Rainer resigns. This causes another decrease in the Survival.
- You see that both events and censored employments decrease the number of employees at risk. But only events cause the estimated survival to change.



### 1.5.3 Code Example

The above results table can be created with the LIFETEST procedure in SAS with the following statements.

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

Note that the TIME statement specifies the two analysis variables.

- DURATION is the variable that holds the length of the time interval for each employee.
- STATUS specifies whether the event was censored or not. In brackets you specify those values that represent censoring events, which is in this case the value '1'.

### Estimating the Average Retention Time

Beside the tabular output in Table 1.2, the LIFETEST procedure also calculates the mean and the median survival.

**Output 1.5: Quartiles and Mean Estimates for the Retention Time**

Quartile Estimates				
Percent	Point Estimate	95% Confidence Interval		
		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

**Note:** The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

From the output you see that:

- The median survival time is 51 months, which is the month when the Survival falls under 0.5.
- The mean survival time in this example is 39.95 months (with a standard error of 5.2).
- If the largest observation is censored and no event time is available, you receive a note that the estimates for the mean survival are underestimated as it had to be restricted to the last observed duration value.

### Interpretation

You can conclude that the mean survival of employees in the SALES\_ENGINEERS department is around 3 years and 4 months (about 39.9 months, as shown in Output 1). Interpreting the median, you can conclude that after 4 years and 3 months (51 months, as shown in Output 1), half of the SALES\_ENGINEERS left the company.

The important difference of these results is that they are not based on arbitrary assumptions about the remaining lifetime of actual employees and no observations are excluded from the analysis.

## 1.5.4 Graphical Representation of the Kaplan-Meier Curve

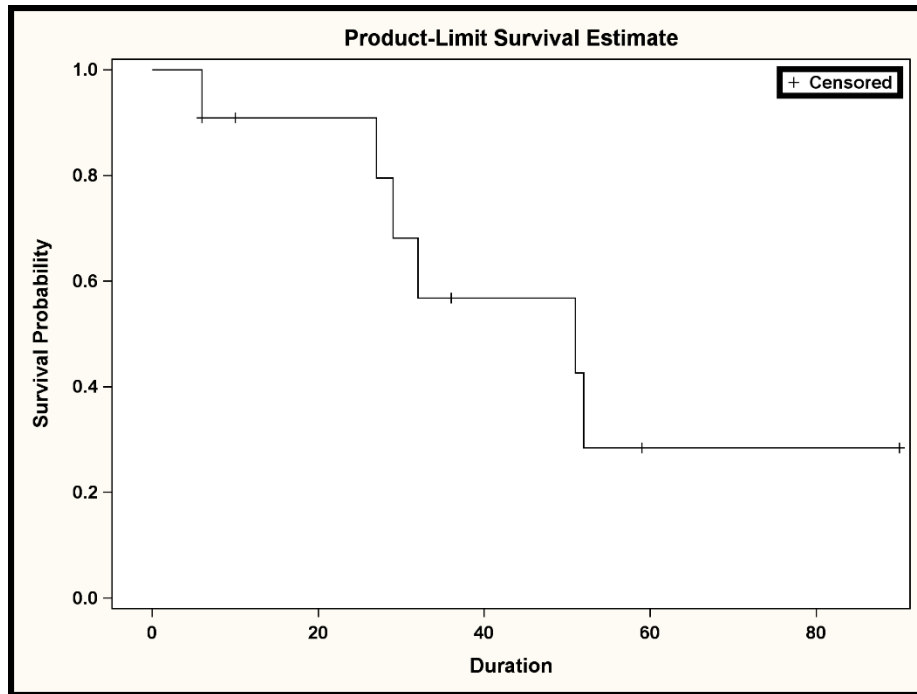
### Graphical Representation

In Figure 1.1 you see the survival curve for the above example. If ODS Graphics are turned on in your SAS session, this chart is automatically created from the LIFETEST procedure call as shown above.

You can turn on ODS Graphics with the following SAS statement:

```
ods graphics on;
```

**Figure 1.1: Survival Curve for the SALES ENGINEERS**



### Interpretation

- You see that the survival curve has the value 1 at the start of the observation period (duration=0).
- The survival curve is a step curve that drops at those time points, when an employee resigns.
- Referencing the data in Table 1.2, you see that the first four steps in the curve are those when John, Rainer, Vincenz, and George resign.
- Employees that are censored from the analysis at a particular point in time are represented with a '+' sign. Here the survival curve does not change its course.
- You see the steps get steeper with increasing duration, accordingly, the hazards increase. This is due to the fact that fewer employees are at risk at that time and one event has a larger effect. The hazard rate quantifies the instantaneous risk that an event occurs at a particular event time. (Compare this to Allison [1], page 16.)
- The last observation (Alan) is censored at month 90. Thus, the survival curve does not drop to 0.
- At the horizontal axis, the number of employees that are still with the company after a certain duration are printed as the "at-risk" population.

## 1.6 Detailed Analysis of the Survival Curve

### 1.6.1 Creating the Survival Curve for All Employees

#### SAS Code

In the previous section only employees from the SALES\_ENGINEER department have been analyzed. If you run the analysis on all employees with the following statement, you will see the output shown in Output 1.6.

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

#### Survival Estimates

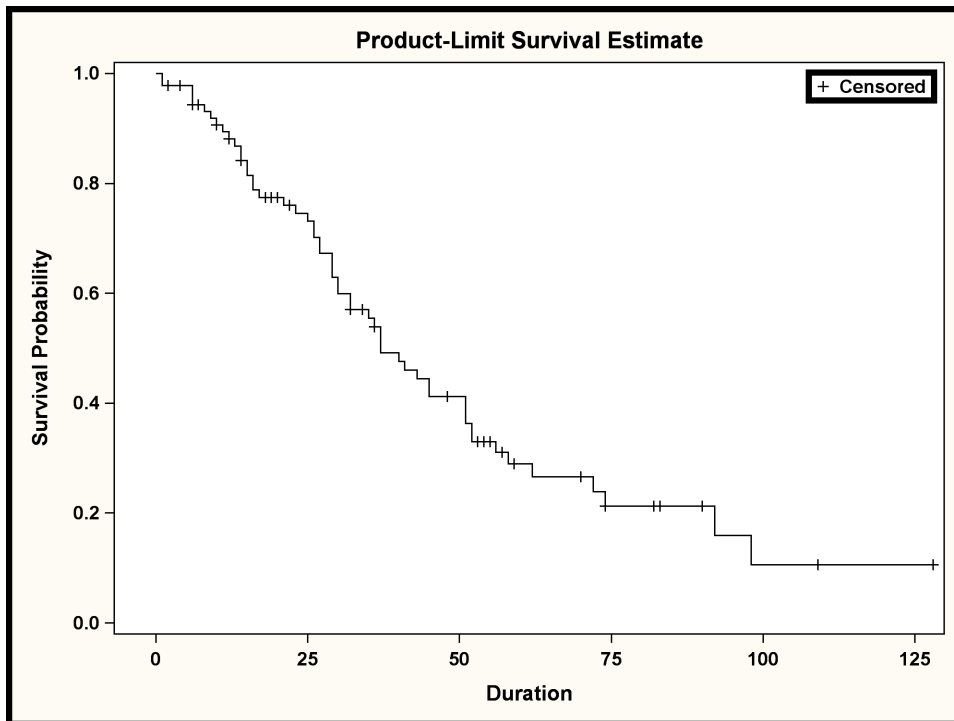
The procedure output contains the product-limit survival estimates, which is partially shown in Output 1.6. This information can be interpreted in the same way as discussed earlier in Table 1.2.

Note that the value for the survival estimate is missing for the censored observations as these records do not indicate any change in the survival. Only records that relate to events change the survival estimate. The survival curve as shown in Figure 1.2 is a step function that only changes for the event records, where a new survival estimate value can be calculated.

**Output 1.6: Screenshot of the Standard Output Objects of the LIFETEST Procedure (Truncated)**

Product-Limit Survival Estimates						
Duration		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000		1.0000	0	0	0	91
1.000		.	.	.	1	90
1.000		0.9780	0.0220	0.0154	2	89
2.000	*	.	.	.	2	88
2.000	*	.	.	.	2	87
4.000	*	.	.	.	2	86
4.000	*	.	.	.	2	85
6.000		.	.	.	3	84
6.000		.	.	.	4	83
6.000		0.9435	0.0565	0.0246	5	82

Figure 1.2: Survival Curve for All Employees



This curve is based on 91 observations. When you compare it to Figure 1.1 that was created only for the sales engineers, you see that there are more and smaller steps and the course of the curve is smoother.

### Average Survival

You also receive the quartile estimates as shown in Output 1.7. The median employee retention time in this company is 37 months with a confidence interval of 30 and 51 months. The estimated mean survival (46.8 months) is a little bit larger than the median.

Output 1.7: Median and Mean Survival and Censoring information

Quartile Estimates				
Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
75	72.000	LOGLOG	51.000	.
50	37.000	LOGLOG	30.000	51.000
25	23.000	LOGLOG	14.000	29.000

Mean	Standard Error
46.757	3.813

**Note:** The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Summary of the Number of Censored and Uncensored Values			
Total	Failed	Censored	Percent Censored
91	54	37	40.66

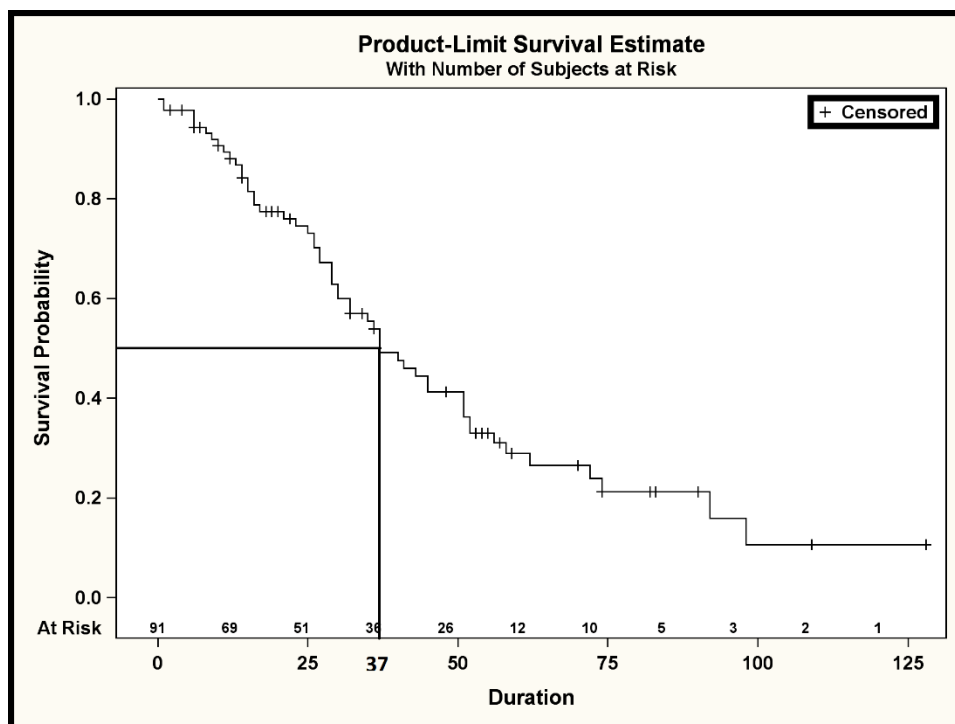
The output also shows that 54 of the 91 observations have an observed end-of-career date, while 37 observations have been censored in the analysis. When this analysis took place in January 2017, those 37 had an active employment with the company.

## 1.6.2 Interpreting the Survival Curve

### Reading from the Survival Curve

In Figure 1.3 you see the survival curve for all employees. The graph allows you to visually identify the median survival by drawing a horizontal line at Survival 0.5 toward the survival curve. The value at the X-axis, 37 months, is the median survival. A bold solid line has been added to the survival curve in Figure 1.3 to illustrate this.

Figure 1.3: Survival Curve for all Employees with Employees at Risk



### Displaying the Population at Risk

The at-risk population decreases on the duration axis from left to right because of two reasons.

- Observations have an “event” and the survival curve drops at these points.
- Observations are censored from the analysis. The occurrence of censored observations is indicated as a ‘+’ in the survival curve.

For better interpretation of the survival curve, the number of analysis subjects at risk is usually printed above the horizontal axis, see also Figure 1.3. It allows you to get an impression of how many observations are used to estimate the survival at different time values.

Above the X-axis the number of employees that are not censored or have not resigned until that time are displayed in 12-month intervals.

In order to display the number of analysis subjects at risk, you need to specify it in the PLOTS= option in the LIFETEST procedure.

```
PROC LIFETEST DATA=employees PLOTS=survival(ATRISK=0 to 120 by 12) ;
  TIME Duration*Status(1);
RUN;
```

As calendar months are considered in the analysis, a BY group of 12 months makes sense. This displays per employment year, the number of employees that are in the analysis.

Note that the creation of the survival plot is the default in the LIFETEST procedure if the ODS GRAPHICS is turned on. Thus, the PLOTS= option has not been specified in the previous examples. If you want however to specify additional options, for example, displaying the number of analysis subjects at risk, you need to explicitly specify it.

---

### 1.6.3 Adding Confidence Bands to the Survival Curve

#### SAS Code

Confidence intervals increase the amount of information that can be retrieved from the results. Displaying these intervals in the graph allows you to assess the certainty of your results.

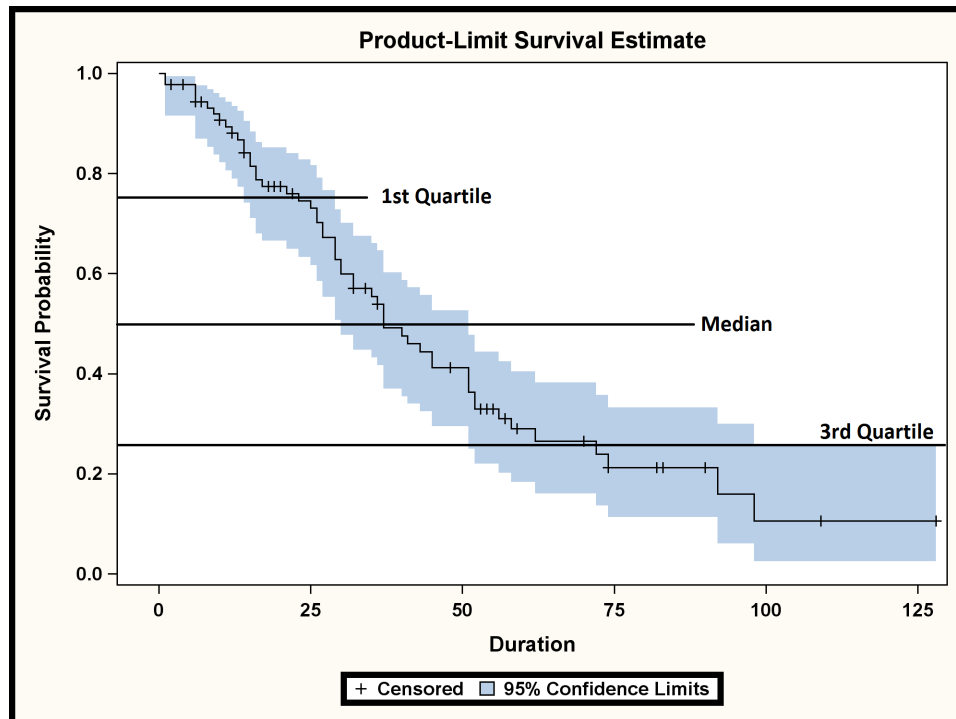
In Output 1.5 the confidence interval of the median survival has already been shown. This confidence band can also be added to the plot of the survival curve by using the following statements.

```
PROC LIFETEST DATA=employees PLOTS=(survival(cb=hw) );
  TIME Duration*Status(1);
RUN;
```

The CB= option requests a confidence band for the survival plot. The value EP specifies the equal precision confidence band. Figure 1.4 shows the output.

## Output and Discussion

Figure 1.4: Survival Curve for All Employees with a 95% Confidence Band



In order to facilitate the reading of the values, black solid lines have been added to the graph. The thick horizontal line at value 0.5 crosses the confidence band at value 30 and at 51. This equals the value for the 95% confidence interval for the median survival in Output 1.5.

Values for the 1<sup>st</sup> quartile at value 0.25 and for the 3<sup>rd</sup> quartile at value 0.75 can be read and compared with Output 1.5. This results in 23 (14-29) and 72 (51-.) respectively. Note that upper limit for the 0.75 quartile cannot be determined, as here the band extends until the end of the observation period.

---

## 1.7 Interpreting the Hazard Curve

### 1.7.1 Basic Idea of the Hazard Curve

The only plot that has been shown so far is the survival curve. This allows you to display the decrease in the number of analysis subjects that are in the analysis over time. In Chapter 2 you will see that this type of visualization is especially useful, when the survival curve between two or more groups shall be compared.

The hazard curve displays the risk over time of an analysis subject to have an event. In the context of the business case study described above, the hazard curve shows the risk of ending an employment over time. This allows a good interpretation of the events and phases in the “lifetime” of an employee and the risk of ending the employment in a particular period.

Chapter 2 in Allison [1] contains a very good discussion on the interpretability of the hazard function and its mathematical definition.

---

### 1.7.2 Adding a Plot for the Hazard Curve

You create a hazard plot as shown in Figure 1.5 with the following statements:

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

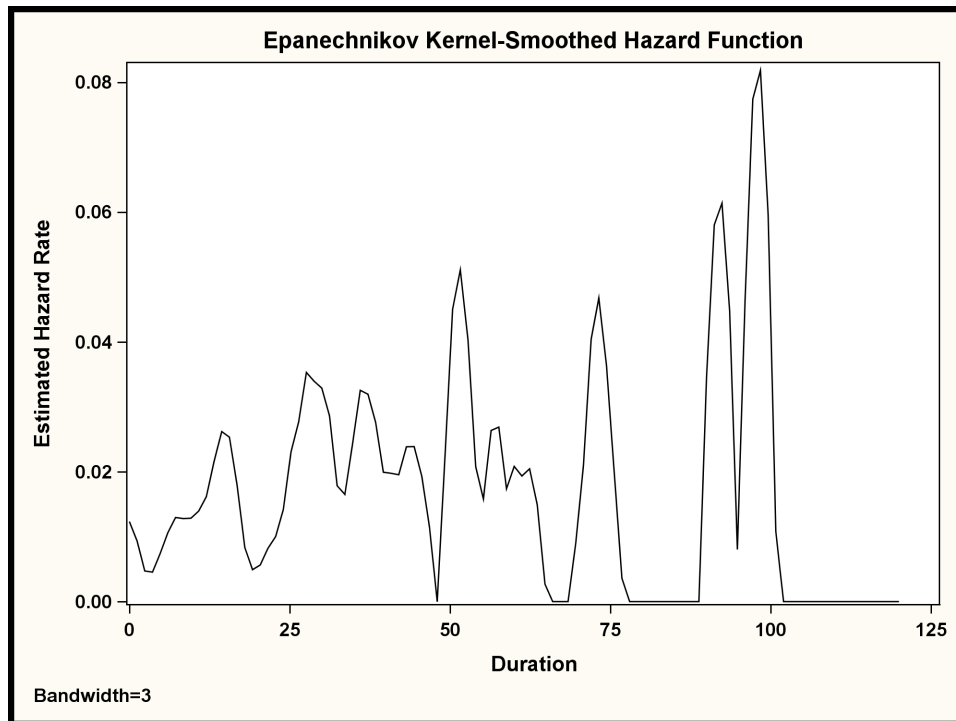
Note that the BANDWIDTH option is important here as it specifies how the hazard rate is smoothed.

Figure 1.5 shows the hazard curve over time for all employees. A kernel smoothing with a bandwidth of 3 months has been used for the display of hazard rate at the Y-axis. The details section in *SAS/STAT® 9.4 User's Guide* [2] contains formulas for finding the optimal bandwidth.

This chart allows you to study the hazard for a resignation at each point in time. You see that the curve is getting more erratic in later time periods. This is due to the lower number of employees at risk here, and one resignation has a higher relative effect.

In the first 2 years, the hazard to resign the job is rather low (except a peak around month 12-15). Then the hazard rate increases until month 60.

**Figure 1.5: Hazard-Curve for All Employees**



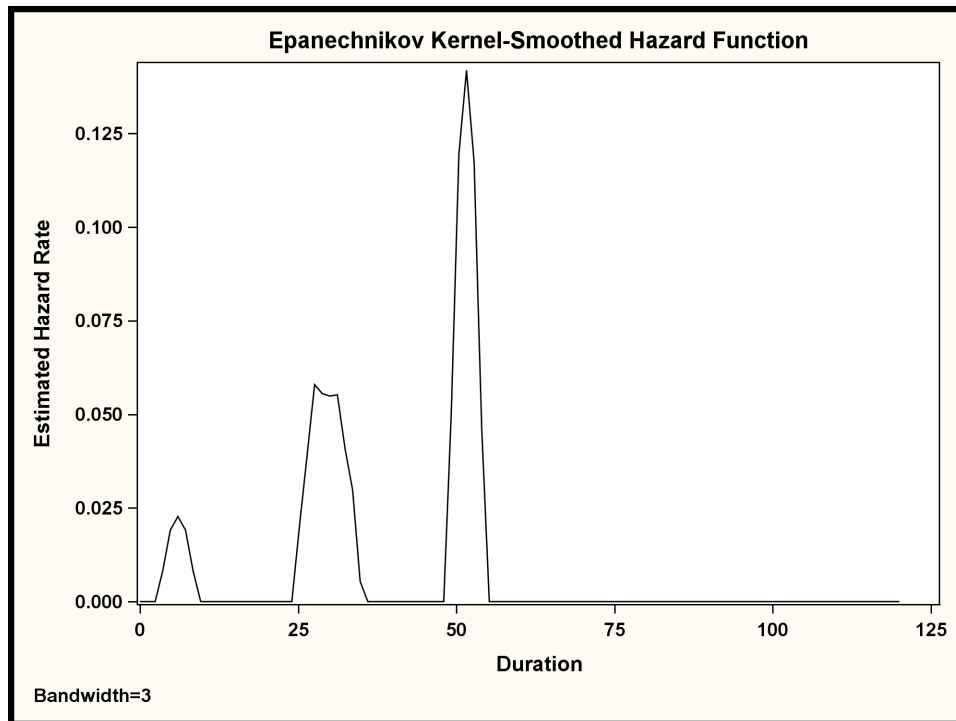
### 1.7.3 The Hazard Curve for the SALES\_ENGINEER Department

#### Creating the Results

The hazard curve in Figure 1.6 for the SALES\_ENGINEER department has been created with the following code:

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



**Figure 1.6: Hazard Curve for the SALES\_ENGINEERS**

### Business Reasoning

The hazard curve in Figure 1.6 gives you an impression about the events taking place over time for the SALES\_ENGINEER department. You can see how resignations distribute over the employees' lifetime and identify three waves based on business assumptions:

- Short-term resignations (after half of a year) of employees that realize that the job does not meet their expectations or that they do not fit to the job.
- Resignations after two years of employment of employees who expected a raise or a senior position at that time.
- Resignations after four years of employment of employees looking for new challenges after that time period.

---

## 1.8 Additional Methods in PROC LIFETEST

### 1.8.1 Using the Lifetable Method

#### General Idea

By default, PROC LIFETEST creates Kaplan-Meier estimates for the survival curve. With that method every individual observation in the input data results in one row in the Kaplan-Meier estimates table. In the case of large data sets with many events, this might cause a long runtime and a very long output file.

An alternative is to use the lifetable method. You specify the option `METHOD = LIFE` to request this analysis. Option `INTERVALS` allows you to specify the intervals that are used for the lifetable calculation. Here you get an output table where every interval is represented by one row. For each interval the number of events and censored observations are shown.

## SAS Code

The following code creates the survival estimate as a lifetable with 6-month intervals.

```
PROC LIFETEST DATA=employees
                METHOD=LIFE INTERVALS=0 to 120 by 6;
    TIME Duration*Status(1);
RUN;
```

## Output Table

Selected columns of the results and rows of the lifetable results are shown in Output 1.8:

- the time intervals into which the failure and censored times are distributed. Each interval is from the lower limit, up to but not including the upper limit; if the upper limit is infinity, the missing value is printed.
- the number of events that occur in the interval
- the number of censored observations that fall into the interval
- the effective sample size for the interval
- the estimate of conditional probability of events (failures) in the interval
- the standard error of the conditional probability estimator
- the estimate of the survival function at the beginning of the interval
- the estimate of the cumulative distribution function of the failure time at the beginning of the interval

Compare the details section in *SAS/STAT® 9.4 User's Guide* [2] for a complete list.

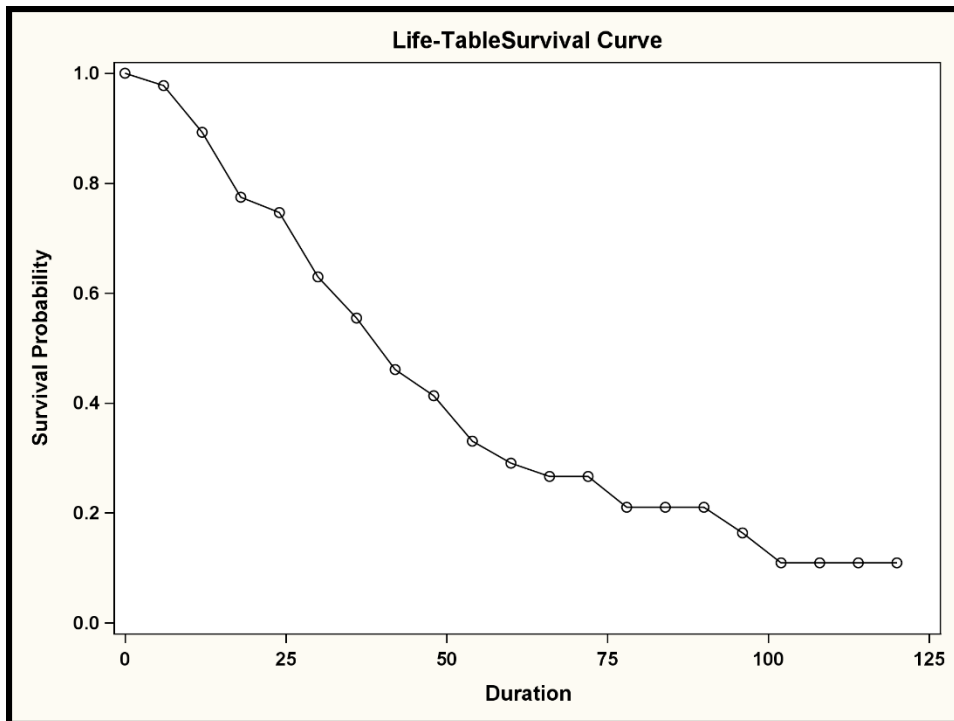
**Output 1.8: Survival Estimates Based on the Lifetable Method (Selected Columns and Rows Only)**

Life Table Survival Estimates									
Interval		Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	
[Lower,	Upper)								
0	6	2	4	89.0	0.0225	0.0157	1.0000	0	
6	12	7	9	80.5	0.0870	0.0314	0.9775	0.0225	
12	18	9	2	68.0	0.1324	0.0411	0.8925	0.1075	
18	24	2	5	55.5	0.0360	0.0250	0.7744	0.2256	
24	30	8	0	51.0	0.1569	0.0509	0.7465	0.2535	
30	36	5	2	42.0	0.1190	0.0500	0.6294	0.3706	
36	42	6	1	35.5	0.1690	0.0629	0.5545	0.4455	
42	48	3	0	29.0	0.1034	0.0566	0.4608	0.5392	
48	54	5	2	25.0	0.2000	0.0800	0.4131	0.5869	
54	60	2	5	16.5	0.1212	0.0803	0.3305	0.6695	
60	66	1	0	12.0	0.0833	0.0798	0.2904	0.7096	

## Survival Plot

The survival curve for the lifetable method can be plotted in the same way as for the Kaplan-Meier method. Depending on the width of the intervals, you end up with a survival curve with a different number of steps.

Figure 1.7: Survival Plot for the Lifetable Method



### 1.8.2 Generating an Output Data Set

Using the OUTSURV= option you can output the survival estimates table to a data set. The following code creates a data set SurvTable as shown in Output 1.9.

```
PROC LIFETEST DATA=employees OUTSURV = SurvTable;
  TIME Duration*Status(1);
RUN;
```

This data set contains one row per analysis subject as presented in the input data. For each observation the duration and the censoring flag is shown. The estimated survival function with the lower and upper confidence limit is shown. This data can be used to create your own customized plots of the survival function.

Output 1.9: Output Data Set Containing the Survival Function

Duration	_CENSOR_	SURVIVAL	SDF_LCL	SDF_UCL
0	.	1	1	1
1	0	0.978021978	0.9149733203	0.9944576201
2	1	0.978021978	.	.
2	1	0.978021978	.	.
4	1	0.978021978	.	.
4	1	0.978021978	.	.
6	0	0.9435035553	0.8695268681	0.9760994457
6	1	0.9435035553	.	.
6	1	0.9435035553	.	.
6	1	0.9435035553	.	.
6	1	0.9435035553	.	.
7	1	0.9435035553	.	.
8	0	0.9312502623	0.853214008	0.9685463496
9	0	0.9189969694	0.8373883515	0.9605896758
10	0	0.9067436765	0.8219400851	0.952301452

## 1.9 Conclusion

This chapter has shown that survival analysis is an excellent tool for analyzing time-to-event data. The Kaplan-Meier method allows you to consider both events and censored observations in the analysis. Different to calculating simple averages and making arbitrary assumptions about the data, this method uses all of the available data for the analysis and allows you to draw conclusions about the average time period. It provides you with a universal method to deal with such information without depending on particular assumptions or losing information or removing analysis subjects from the data.

While the method is widely used in medical statistics and event time analyses in engineering, the case study has shown that it provides valuable insight in other domains as well. Investigating survival curves or hazard curves shows you how different events or phases in the individual life time relate to different courses in survival.

The survival plot and the hazard plot give a visual impression about the course over time and allow an interpretation from a business point of view.

So far the analyses have only been performed for a single group. The next chapter reveals even more power of the survival analysis method, when different groups are compared.

### Coding

SAS code for the LIFETEST procedure has been shown to run these analyses.

### Performance Considerations and Scalability

In the default setting, the LIFETEST procedure uses the Kaplan-Meier method for the analysis. With that method every individual observation in the input data results in one row in the Kaplan-Meier estimates table. In the case of large data sets with many events, this might cause a long runtime and a very long output file.

An alternative is to use the lifetable method as shown in Section 1.8.1.

# About This Book

---

## Rationale to Write This Book

---

### In a Nutshell

This book reflects my enthusiasm to use analytical and data science methods to solve business questions and to implement the solution using SAS.

---

### Importance of Analytical Methods

#### More Than Descriptive Statistics

Over the last few years I answered many business questions from our customers using analytical methods. For most of these questions, the application of analytical methods made a large difference. It allowed me to cover the business questions in a much more comprehensive, precise, and detailed way compared to the application of only graphical or descriptive methods.

That experience led me to write a book that illustrates and explains how data science methods and analytical methods can be applied in different business domains for different business questions.

#### Business Focus

The idea of this book is to provide a collection of case studies that have a business relationship and that show that analytical methods contribute value.

I also learned that many methods that are well established in a certain industry or business domain can be beneficial for other areas as well. One example is the application of survival analysis methods for employee headcount analysis, as shown in the first case study.

---

## The Power of the SAS Analytics Platform for Analytics and Data Science

The first SAS program that I wrote dates back to 1991. Over the years I had the pleasure to combine my analytical knowledge and the power of the SAS Analytics Platform in many projects across different industries and business domains.

In these projects I experienced the importance of being able to combine advanced analytics with data management and reporting capabilities. It is one of the key paradigms of the SAS Analytics Platform to seamlessly provide this functionality.

This allows me, as a data scientist, to perform all my analysis tasks in a single environment.

- Preparing data
- Checking and improving data quality
- Applying data science methods and generating results
- Preparing and enhancing the results for further, more detailed analysis
- Presenting the results in a format that is appropriate for the information consumer

My first two books [6 and 9] have a focus on the first two bullet points of this list. This book follows a case study approach and focuses on the application of data science methods, and the preparation, enhancement, and presentation of the results.

It illustrates the perfect fit of using the SAS Analytics Platform for the analysis of various business questions with data science methods.

---

## Who Should Read This Book

This book is written for a variety of different persona groups and profiles.

### Business Analysts and Business Experts

Businesspeople can review the examples and see what can be achieved with analytical methods. They get insight into the power of analytics and the additional findings that can be generated by these methods. They might not study the SAS implementation and the code in much detail. They would rather hand over the implementation examples to their data scientist to give them a quick start to apply the methods.

### Statisticians, Data Miners, Data Scientists, and Quantitative Experts

This group of people might be interested to see how analytical methods can be applied to real-world business questions. They learn how analytical methods that are established in a certain industry might be applied to other areas. They see practical situations and constraints that they can expect to encounter when they apply data science methods.

### SAS Programmers

The book contains a lot of SAS code, including SAS macros, SAS DATA step code for data preparation, SAS analytics procedures, and SAS graph procedures. In this code SAS programmers can find new ways to solve certain problems in SAS and transfer the solutions in these examples to their day-to-day problems.

---

## Data Science Methods and SAS Procedures Covered in this Book

---

### This Book Covers the Following Data Science Methods

- Kaplan-Meier Estimates – Cox Proportional Hazards Regression – Survival Data Mining
- Smoothing of Longitudinal Data – Multivariate Adaptive Regression Splines – Automatic Breakpoint Detection – Automatic Detection of Outliers – ARIMA Models
- Linear Regression – Poisson Regression – Quantile Regression – New Product Forecasting – Similarity Search
- Imputation of Missing Values – Association Analysis – Benford's Law – Chi2 Independency Test
- Monte Carlo Simulation – Mathematical Programming – Data Matrices – Simulation of Complex Processes

---

### The Following SAS Procedures and SAS Solutions Are Used

#### Analytic SAS Procedures

LIFETEST – PHREG – ARIMA – X11 – X13 – ADAPTIVEREG – HPFDIAGNOSE – VARCLUS –  
TREE – HPGENSELECT – GLMSELECT – GLM – QUANTSELECT – QUANTREG –  
HPQUANTSELECT – IML

#### Data Management and Graphical Procedures Graphic

SGPLOT – SGPANEL – GTILE – FREQ – MEANS – TRANSPOSE – SQL  
SAS DATA step, SAS Macro Language

#### SAS Enterprise Miner

Survival node  
Association node

## Structure of This Book

### Overview of the Case Studies

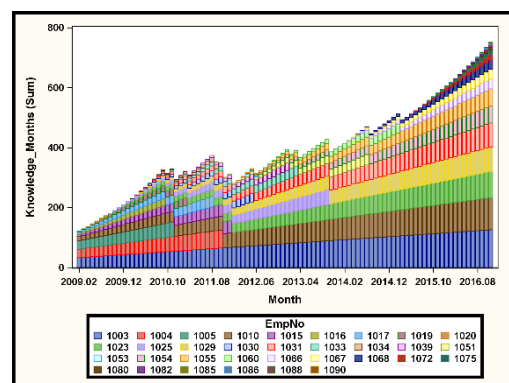
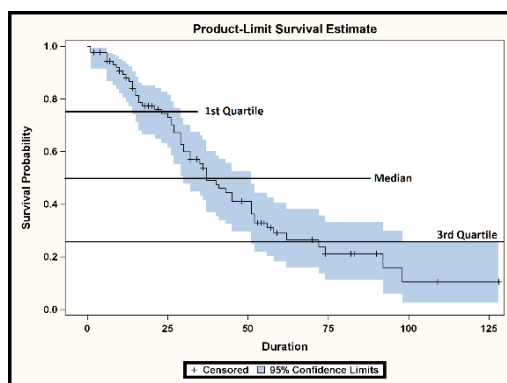
This book covers 8 case studies in 28 chapters. This section gives an overview of the case studies, the chapters, the rationale, the main business questions, and the analytical methods that are used to answer these questions.

#### Case Study 1 – Performing Headcount Survival Analysis for Employee Retention

This case study uses employee retention data to illustrate how analytical methods allow you to draw conclusions about the average length of time intervals, even if most of the endpoints have not yet been observed.

Survival analysis methods like Kaplan-Meier estimates and Cox Proportional Hazards regression are used to solve the business questions. The case study contains the following chapters:

1. Using Survival Analysis Methods to Analyze Employee Retention Time
2. Analyzing the Effect of Influential Factors on Employee Retention Time
3. Performing Survival Data Mining - The Data Mining Approach for Survival Analysis
4. Visualizing Employee Retention Data

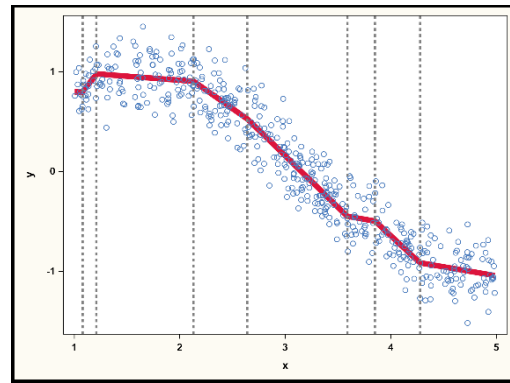
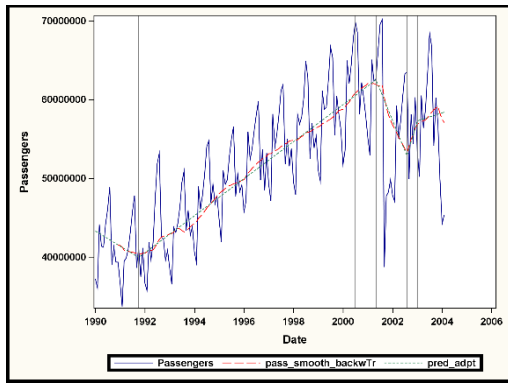


#### Case Study 2 – Detecting Structural Changes and Outliers in Longitudinal Data

This case study shows how analytical methods can be used to automatically detect events and changes in the course of longitudinal data. Example time series data with the number of airline passengers and data from a long-term clinical trial are used to illustrate how data can be smoothed and breakpoints and outliers can be detected.

Analytical methods like multivariate adaptive splines regression, ARIMA models, and moving averages are used to solve the business questions in the following chapters:

5. Analyzing and Smoothing the Course of Longitudinal Data
6. Detecting Structural Changes in Longitudinal Data
7. Detecting Outliers and Level Shifts in Longitudinal Data
8. Results from a Simulation Study with Longitudinal Data
9. Analyzing the Variability of Longitudinal Data



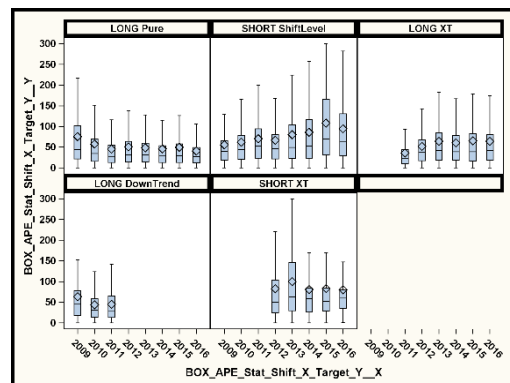
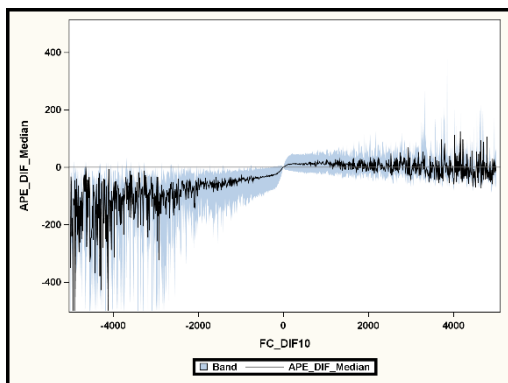
### Case Study 3 – Explaining Forecast Errors and Deviations

This case uses regression methods to identify influential factors that have an impact on the forecast accuracy of time series forecasting models. The forecast error usually differs between factors like product group, forecast horizons, and the analytical method that was used to create the forecast. Analytical methods allow you to identify and isolate these effects to provide more insight into the generation of forecasts.

This case study also deals with the important question of whether demand planners really improve forecast accuracy with their manual overrides of the statistical forecast.

Linear regression and quantile regression are used to analyze these questions in the following chapters:

10. Investigating Forecast Errors with Descriptive Statistics
11. Investigating Forecast Errors with General Linear Models
12. Interpreting the Coefficients of Categorical Variables in Regression Models
13. Using Quantile Regression to Get More Than the Average Picture
14. Analyzing the Effect of Manual Overrides in Forecasting



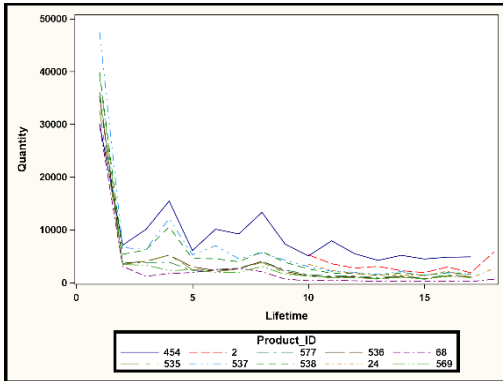
### Case Study 4 – Forecasting the Demand for New Products

This case study shows how demand forecasts can be generated for products that have no or only a short time history of known demand.

Methods like Poisson regression or similarity search are used to solve this business question in the following chapters:

15. Performing Demand Forecasting for New Products
16. Using Poisson Regression to Forecast the Demand for New Products
17. Using Similarity Search to Forecast the Demand for New Products



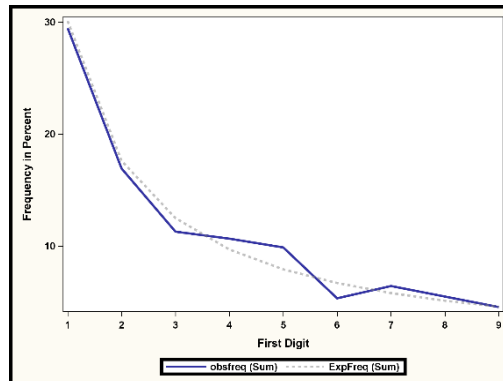
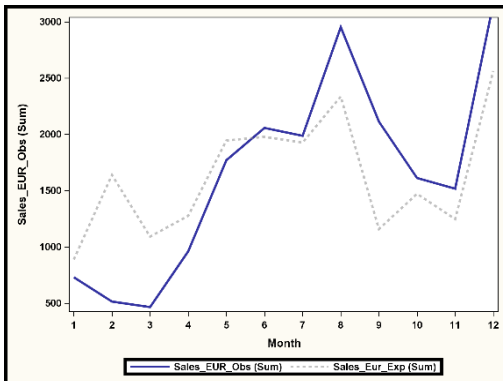


### Case Study 5 – Checking the Alignment with Predefined Patterns

A frequent business question is to verify whether different entities our counterparts show the expected behavior or adhere to predefined patterns or processes. For the different interaction strategies you want, for example, to know which customers show a behavior that is far from what you expected. In financial accounting, the analysis of Benford’s law is often investigated.

Methods like the  $\chi^2$  independency test are used to verify these assumptions in the following chapters:

- 18. Checking Accounting Data for the Benford’s Law
- 19. Checking the Benford’s Law for Multiple Accounts
- 20. Checking Different Patterns in the Data

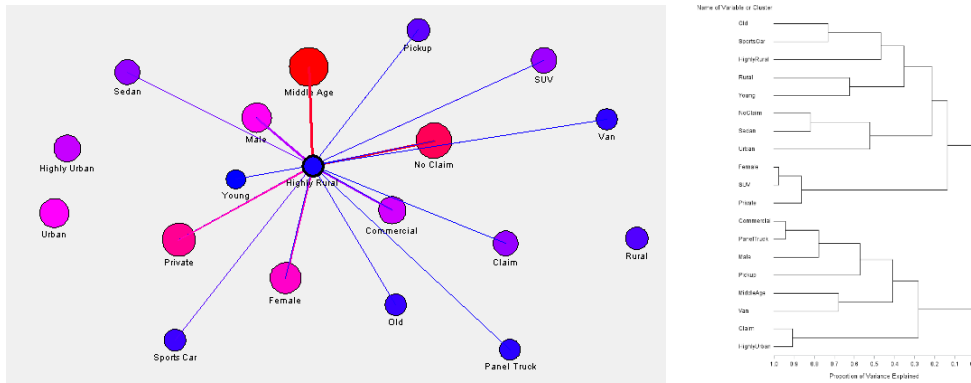


### Case Study 6 – Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

This case study shows how you can receive answers from your data, even if you do not ask every question in detail. You see which features and properties in the data are closely related together.

Unsupervised machine learning methods like association analysis and variable clustering are used in the following chapters:

- 21. Finding Relationships in Your Analysis Data with Association Analysis
- 22. Using Variable Clustering to Detect Relationships in Your Data
- 23. Investigating of Clinical Trial Data in an Explanatory Way

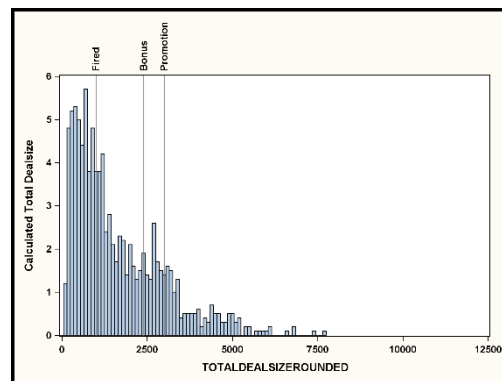
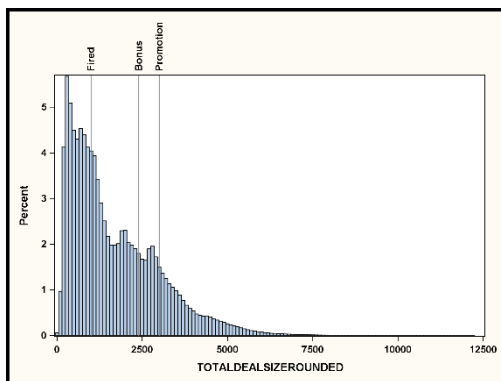


### Case Study 7 – Using Monte Carlo Simulations to Understand the Outcome Distribution

This case study shows how simulation studies can be used to get a more comprehensive picture about the outcome distribution. The case study uses the sales projects pipeline of a sales manager and answers the questions about the likelihood that the sales manager might get fired because he misses a certain minimum target.

Methods like Monte Carlo simulations are used in these chapters. An approach using matrix calculations with SAS/IML software is shown.

- 24. Calculating the Outcomes of All Possible Scenarios
- 25. Using Monte Carlo Methods to Simulate the Distribution of the Outcomes

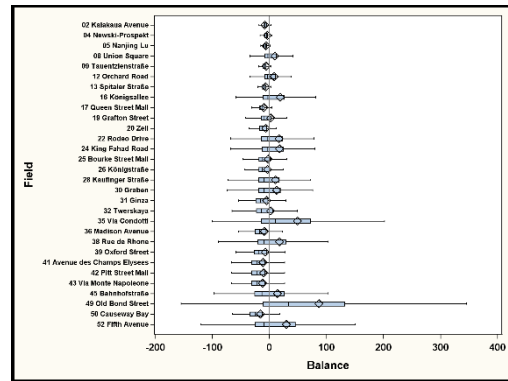
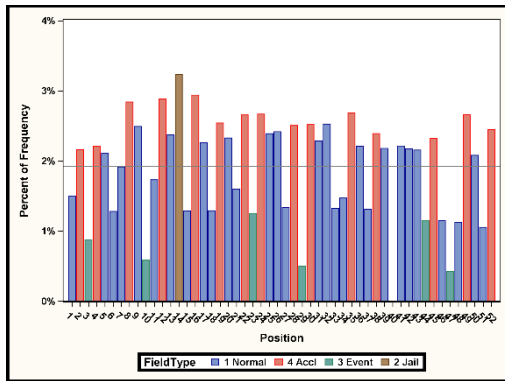


### Case Study 8 – Studying Complex Systems – Simulating the Monopoly Board Game

Learning more details about the behavior of complex systems and the relationship between different components of this system is very often needed. This case study shows how Monte Carlo simulations can be used to simulate the Monopoly board game. The simulations include analysis of the visit frequency on different fields of the board game as well as a profitability analysis of different properties.

Monte Carlo simulations with a SAS DATA step are shown in this case study.

- 26. Creating a Basic Framework to Simulate the Visit Frequency on the Fields of the Monopoly Board Game
- 27. Enhancing the Simulation Framework to Consider Special Rules
- 28. Simulating the Profitability of the Property Fields of the Monopoly Board Game




---

## Importance of Advanced Analytics and Data Science Methods

---

### Introduction

Some relationships in the data can also be spotted graphically or with simple descriptive methods. Advanced analytics and data science methods are very important as they provide insight on a more detailed level into the nature and the extent of different relationships. Note that in the list below, the terms *analytical methods* and *data science methods* are used synonymously.

---

### Generating Results Automatically

Descriptive or interactive analyses often require that you manually perform the analysis for each group. They also require that you explicitly search for every feature in the data. Analytical methods, however, allow you to identify relationships automatically by specifying, for example, a list of influential factors. These methods automatically select and assess the most relevant factors. This can be performed even for a large number of input features and subgroups.

---

### Being More Objective

Analytical methods are objective in their results. They are not influenced by personal preferences or intuition. The results are solely based on the facts that can be seen in the data and that can be formulated into mathematical relationships. The results are thus not influenced by personal opinions, individual experiences, and events in the past that are differently assessed by different people.

---

### Eliminating the “Yes, but ...” Phrases

Analytics methods consider environmental parameters directly into the model. Seasonal variation or the fact that some regions show a different outcome can be included into the model. The forecasts and predictions of such models are thus automatically corrected for these side effects. The process of defining decisions usually eliminates phrases such as the following:

- Yes, but in this region we usually have a higher response.
- Yes, but in the winter season we have to expect lower demand.

This knowledge and the objective impact of these events and side conditions have already been considered in a model.

---

### Handling Multivariate Relationships

While humans are good at making intuitive decisions and can handle univariate influential factors like “Younger customers have a higher shopping frequency”, we often fail to consider multivariate relationships in an appropriate way. Analytical methods can be trained to consider multivariate relationships and also interactions between different factors and can thus provide more detailed results and deeper insights.

---

## Handling Complex Situations

Analytical methods can analyze questions that cannot be solved with simple descriptive statistics. You might also be able to identify some results graphically, like the average influence of customer demographics on the response probability of customers.

There are, however, many situations where you need data science methods to receive information, like the case of time to event data, where the true endpoint is unknown for many analysis objects and no business rule for the treatment of such situations is available. This case is explained in the example of employee headcount analysis shown in case study 1.

---

## Quantifying Relationships

Analytical methods not only identify differences between groups and time intervals. The relationships are also quantified in an assessment scheme. Such a scheme allows you to assign probabilities for events or an expected outcome value to other analysis subjects.

## Prioritization of Analysis Subjects for Further Actions

The probabilities and expected outcome values that are retrieved from analytical models can be used to prioritize analysis subjects to be selected for special treatment actions. Analytical methods allow allocating the appropriate attention and resources to those analysis subjects that should be handled first.

## Providing Ranges and Confidence Intervals for the Outcome

In many analyses you are not only interested in the average outcome value but you would like to learn about the most likely range of the outcomes. It makes a difference whether 95 % of the possible outcomes are located in a small value range or whether they are spread over a large value range. Analytical methods provide confidence intervals for both the estimated outcome and the estimated influence of the explanatory factors. This allows you to analyze the nature of the relationships in more detail.

---

## Performing Simulations

Analytical methods allow you to perform simulations and to assess the likelihood of different outcome scenarios. Some systems are too complex to be described by a mathematical model or by simple statistics. Monte Carlo simulations allow you to adjust input parameters and formulate different assumptions about the environment and to see their effect on the outcome. Performing such what-if scenarios allows you to make better decisions.

---

## Downloads and References

For downloads of SAS programs, sample data, and macros that are presented in this book, as well as updates on findings for the different case studies, please visit:  
[http://www.sascommunity.org/wiki/Applying\\_Data\\_Science\\_-\\_Business\\_Case\\_Studies\\_Using\\_SAS](http://www.sascommunity.org/wiki/Applying_Data_Science_-_Business_Case_Studies_Using_SAS).

This site also includes downloadable color versions of selected graphs and figures that are presented in this book. Graphs and figures that reveal their content much better in color are available.

Please visit this site regularly. The author will keep this site up to date and provide updates on the content presented in this book.

In addition, please also see the SAS author page for Gerhard Svolba at  
<http://support.sas.com/publishing/authors/svolba.html>.

The reference section at the end of this book provides suggestions for further reading and more details. The numbers in brackets throughout the text refer to these reference entries.

---

## SAS Environment

### General Comments

The analysis for the case studies has been performed using the fourth maintenance release of SAS 9.4 on a Windows 7 workstation. All programs can be downloaded from the companion site for this book or from [http://www.sascommunity.org/wiki/Applying\\_Data\\_Science\\_-\\_Business\\_Case\\_Studies\\_Using\\_SAS](http://www.sascommunity.org/wiki/Applying_Data_Science_-_Business_Case_Studies_Using_SAS).

### Using the Programs

The programs are provided as one file per chapter and contain the entire code from the book. An additional file contains all the macros that are available with this book.

### Startup Options

The following options were used in the SAS environment.

```
options fmtsearch=(work sasuser) nofmterr validvarname=v7;
ods graphics on;
```

### Using the Data

The programs assume that the example data is available in the WORK library of your SAS session. Once you have downloaded the ZIP file with the data and extracted it to your local drive, you can use the following code to copy the data to the WORK library.

```
libname datasci "c:\tmp\ApplyingDataScience_Data";

proc copy in=datasci out=work;
run;
```

Note that this code assumes that you extracted the data to the library c:\tmp\ApplyingDataScience\_Data. Change the library accordingly to your individual environment.

### Adaption to the Content of the Data Sets

For privacy and security reasons some of the data was changed before it could be provided to the public. This means that using the programs and the data you can technically re-run all the analyses as shown in the chapters. Some of the analyses might, however, produce different results as they are run on adapted data sets.

This is the case for the following datasets:

- MANFC, STATFC, and MATERIAL in case study 3.
- PRODUCT\_BASE and PRODUCT\_DEMAND in case study 4.
- PATIENTS\_MART\_TMP and PATIENTS\_XT in case study 6.

---

## Full SAS Code

In most of the chapters of this book, the presentation of the SAS code is interrupted by textual explanations. This method has been chosen to provide more details for the rationale and the background of different coding options.

Some of the reviewers suggested including the full code of these programs to allow better reading. The code for these programs is presented in Appendix A.

---

## Additional Help

Although this book illustrates many analyses regularly performed in businesses across industries, questions specific to your aims and issues may arise. To fully support you, SAS Institute and SAS Press offer you the following help resources:

- For questions about topics covered in this book, contact the author through SAS Press:
  - Send questions by email to [saspress@sas.com](mailto:saspress@sas.com); include the book title in your correspondence.
  - Submit feedback on the author's page at [http://support.sas.com/author\\_feedback](http://support.sas.com/author_feedback).
- For questions about topics in or beyond the scope of this book, post queries to the relevant SAS Support Communities at <https://communities.sas.com/welcome>.
- SAS Institute maintains a comprehensive website with up-to-date information. One page that is particularly useful to both the novice and the seasoned SAS user is its Knowledge Base. Search for relevant notes in the "Samples and SAS Notes" section of the Knowledge Base at <http://support.sas.com/resources>.
- Registered SAS users or their organizations can access SAS Customer Support at <http://support.sas.com>. Here you can pose specific questions to SAS Customer Support; under Support, click Submit a Problem. You will need to provide an email address to which replies can be sent, identify your organization, and provide a customer site number or license information. This information can be found in your SAS logs.

---

## Keep in Touch

We look forward to hearing from you. We invite questions, comments, and concerns. If you want to contact us about a specific book, please include the book title in your correspondence.

### Contact the Author through SAS Press

- By e-mail: [saspress@sas.com](mailto:saspress@sas.com)
- Via the Web: [http://support.sas.com/author\\_feedback](http://support.sas.com/author_feedback)

### Purchase SAS Books

For a complete list of books available through SAS, visit [sas.com/store/books](http://sas.com/store/books).

- Phone: 1-800-727-0025
- E-mail: [sasbook@sas.com](mailto:sasbook@sas.com)

### Subscribe to the SAS Learning Report

Receive up-to-date information about SAS training, certification, and publications via email by subscribing to the SAS Learning Report monthly eNewsletter. Read the archives and subscribe today at <http://support.sas.com/community/newsletters/training!>

### Publish with SAS

SAS is recruiting authors! Are you interested in writing a book? Visit <http://support.sas.com/saspress> for more information.

## About the Author



Gerhard Svolba was born in Vienna, Austria, in 1970. He studied business informatics and statistics at the University of Vienna and Technical University of Vienna and holds a master's degree. From 1995 until 1999, he was assistant professor in the department for medical statistics at the University of Vienna, where he completed his PhD on statistical quality control in clinical trials (the respective book is published in *Facultas*). In 1999 Gerhard joined SAS Institute Inc. and is currently responsible for the analytical projects in SAS Austria as well as the analytical products and solutions SAS offers.

In 2003, on his way to a customer site to consult with them on data mining and data preparation, he had the idea to summarize his experience in written form. In 2004 he began work on *Data Preparation for Analytics Using SAS*, which was released by SAS Press in 2006. Since then he has spoken at numerous conferences on data preparation and teaches his class “Building Analytics Data Marts” at many locations. In 2012 his next book *Data Quality for Analytics Using SAS* was published. He likes to be in touch with customers and exchange ideas about analytics, data preparation, and data quality.

Gerhard Svolba is the father of three teenaged sons and loves to spend time with them. He likes to be out in nature, in the woods, mountains, and especially on the water, as he is an enthusiastic sailor.

Gerhard Svolba's current website can be found at [http://www.sascommunity.org/wiki/Gerhard\\_Svolba](http://www.sascommunity.org/wiki/Gerhard_Svolba). He answers emails under [sastools.by.gerhard@gmx.net](mailto:sastools.by.gerhard@gmx.net).