# Datenbereinigung und Vorbereitung in JMP 12

Bernd Heinen
SAS Institute GmbH
In der Neckarhelle 162
Heidelberg
Bernd.heinen@jmp.com

**Zusammenfassung**

Die Qualitätssicherung und Aufbereitung von Daten nimmt leicht mehr Zeit in Anspruch, als die eigentliche Analyse. Außerdem ist es eine Tätigkeit, die sich – gelinde gesagt – keiner großen Beliebtheit erfreut. So notwendig diese Arbeit ist, so wenig wird sie honoriert. Falsch eingegebene Begriffe zu finden und zu korrigieren, Ausreißer identifizieren, fehlende Werte erkennen und ersetzen, sind üblicherweise umständlich abzuwickeln. JMP 12 kommt mit einigen kleinen Hilfsfunktionen, die diese Arbeit wesentlich erleichtern.

**Schlüsselwörter:** Ausreißer, fehlende Werte, Rekodierung, JMP12

## 1 Einleitung

Wenn man mit neu erhaltenen Daten arbeitet, sind die Qualitätssicherung und Aufbereitung die ersten und wichtigsten Aufgaben. Üblicherweise bedient man sich dazu derjenigen statistischen Verfahren, die man später auch in der Analyse anwendet. JMP 12 bietet eigene Funktionen an, die statistische Verfahren nutzen und kombinieren, aber in einem eigenen Benutzerdialog zur Verfügung stellen. So bekommt das Spalten Menü einen neuen Wert.

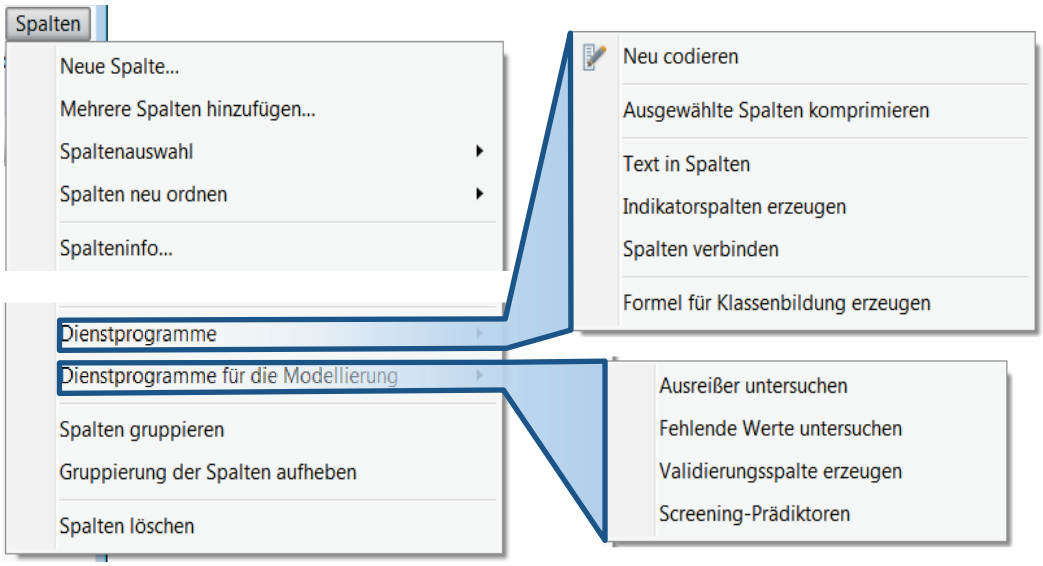



**Abbildung 1:** Neue Funktionen

# 2 Ein Überblick

Eigentlich nicht im Zentrum dieser Themen aber doch am Anfang steht der Hinweis auf die Funktion Spaltenansicht. Diese bietet einen komprimierten Überblick über die Inhalte einer Tabelle. Zahl fehlender Werte für alle Spalten, Zahl der Klassen für kategoriale Größen, Mittelwert, Standardabweichung, Minimum und Maximum für stetige Größen, mehr nicht. Aber schon ein wichtiger erster Schritt für die Einschätzung der Datenqualität. Und da in JMP Berichten jede tabellarische Darstellung in eine eigene Tabelle überführt werden kann (Abbildung 2), lassen sich dazu auch beliebige grafische Darstellungen aufbauen. Mit dem lokalen Datenfilter kann man dann auch unterschiedliche Variablengruppen vergleichen.

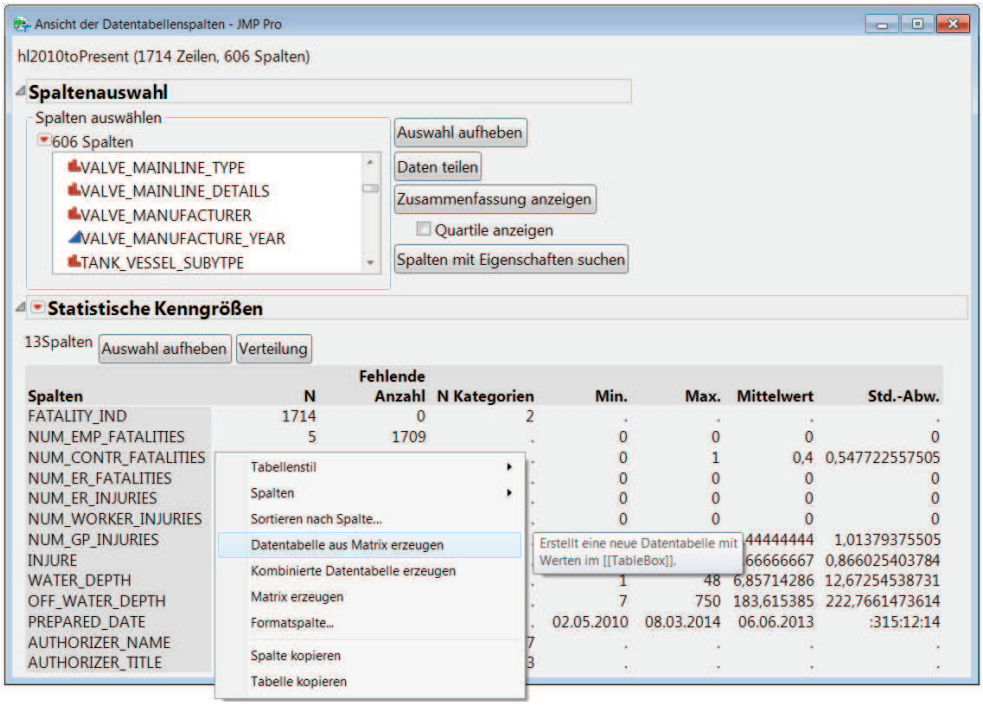



**Abbildung 2:** Spaltenansicht

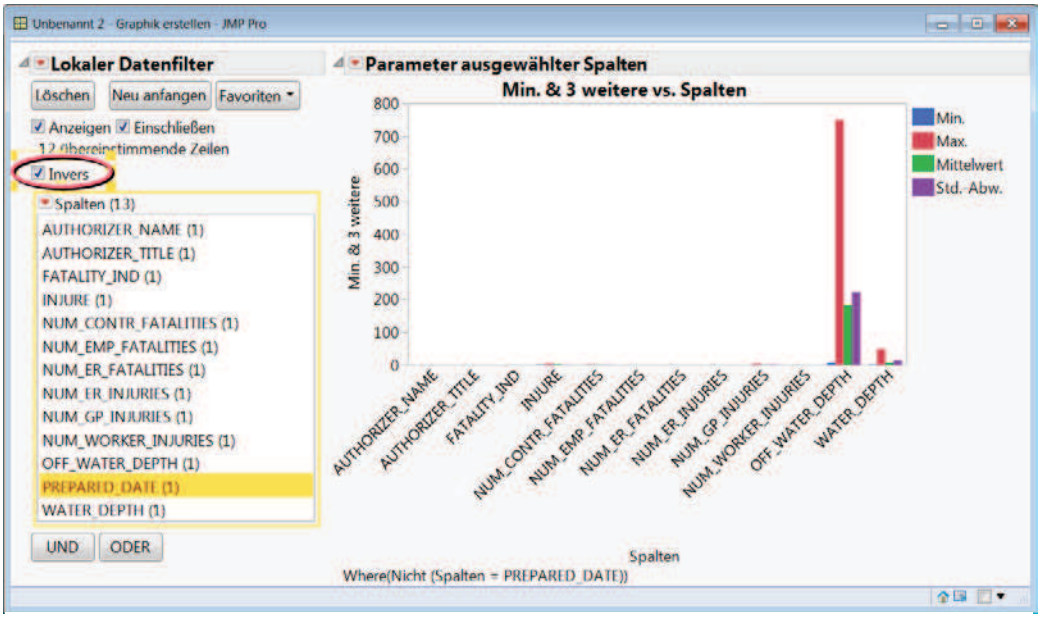



**Abbildung 3:** Spaltenkennzahlen und Datenfilter

# 3 Rekodieren

Trotz weit verbreiteter elektronisch unterstützter Datenerfassung bieten Textspalten immer wieder Überraschungen. Vor allem, wenn viele unterschiedlich Einträge möglich sind, nicht lediglich drei oder vier verschiedene Codes. Einen Eindruck darüber zu gewinnen, was dort eingegeben worden ist, welche unterschiedlichen Einträge aber dasselbe meinen könnten, Leerzeichen zu ignorieren und vieles mehr, ist eine aufwändige Tätigkeit. JMP bietet dazu jetzt eine ganze Anwendung zum neu kodieren von Spalten. Dazu muss eine Spalte ausgewählt sein. Zu Beginn steht eine zweispaltige Tabelle, die


**Abbildung 4:** Kodierungsliste

die Werte alphabetisch auflistet und die jeweiligen Häufigkeiten angibt. Zu den verfügbaren Funktionen gehört das Standardisieren von Groß-/Kleinschreibung, das Reduzieren von Leerzeichen und das Unterteilen von Texten. Sehr mächtig ist die Rekodierfunktion. Sie vergleicht Strings und berechnet deren Ähnlichkeit basierend auf der Edit-Distanz, einer Zahl die angibt, wie viele einzelne Buchstabenänderungen man mindestens benötigt, um ein Wort in ein anderes zu überführen. Man kann die Grenze festlegen, bis zu der Zeichenketten als gleich angesehen werden. JMP gruppiert die Daten neu, zeigt alle Zusammenfassungen an und erlaubt eine manuelle Überarbeitung. Nach jeder Änderung wird angezeigt, auf wie viele eindeutige Werte die Liste reduziert wurde (Abbildung 4).


**Abbildung 5:** Dokumentation der Änderungen

Besonders hilfreich ist, dass zusätzlich viele

Möglichkeiten bestehen, die Ergebnisse und die Vorgehensweise zu speichern (Abbildung 5). So kann man nach erfolgter Bereinigung die Daten in der Spalte selbst ersetzen, man kann eine neue Spalte mit den ersetzten Werten einfügen oder eine Formelspalte einfügen, die die Änderungen als Formel enthält. Ebenso können die Änderungen als Skript gespeichert werden. Skripte können importiert und sogar kombiniert werden, so dass man unterbrochene Arbeit jederzeit wieder aufnehmen kann und am Ende über ein Skript verfügt, das alle Änderungen nacharbeiten und auf neue Daten anwenden kann.

# 4 Indikatoren

Eine häufige Form der Umstrukturierung ist es, Indikatorspalten zu bilden. Meistens keine intellektuelle Herausforderung, oft aber mühsame Arbeit. Hier helfen Funktionen, die die Ausprägungen einer Variablen oder sogar Inhalte einer Zelle analysieren und die entsprechenden Indikatoren erzeugen. Das machen zwei Funktionen: „Text in Spalten“ und „Indikatorspalten erzeugen“.

## 4.1 Text in Spalten

Diese Funktion kann Inhalte einer Zelle nach Position der enthaltenen Worte (1. Wort, 2. Wort, …) aufspalten oder nach den Inhalten (Hund, Katze, Maus,…). Eingeben muss man ein oder mehrere Trennzeichen, danach wird die Funktion ausgeführt. Die Resultate der wortweisen Trennung zeigt die Verteilung der Inhalte auf die verschiedenen Positionen im Gesamttext (Abbildung 6). Durch den Datenfilter sind sieben Namen ausgesucht, an der ersten Position gibt es aber nur fünf verschiedenen Ausprägungen. Auf der zweiten Position gibt es auch sieben Ausprägungen, einen dritten Textbestandteil haben nur vier der ausgewählten Fälle. Es gibt so viele neue Spalten wie maximale Textbestandteile gefunden wurden, die Inhalte der Zellen bestehen aus den jeweiligen Textbestandteilen.

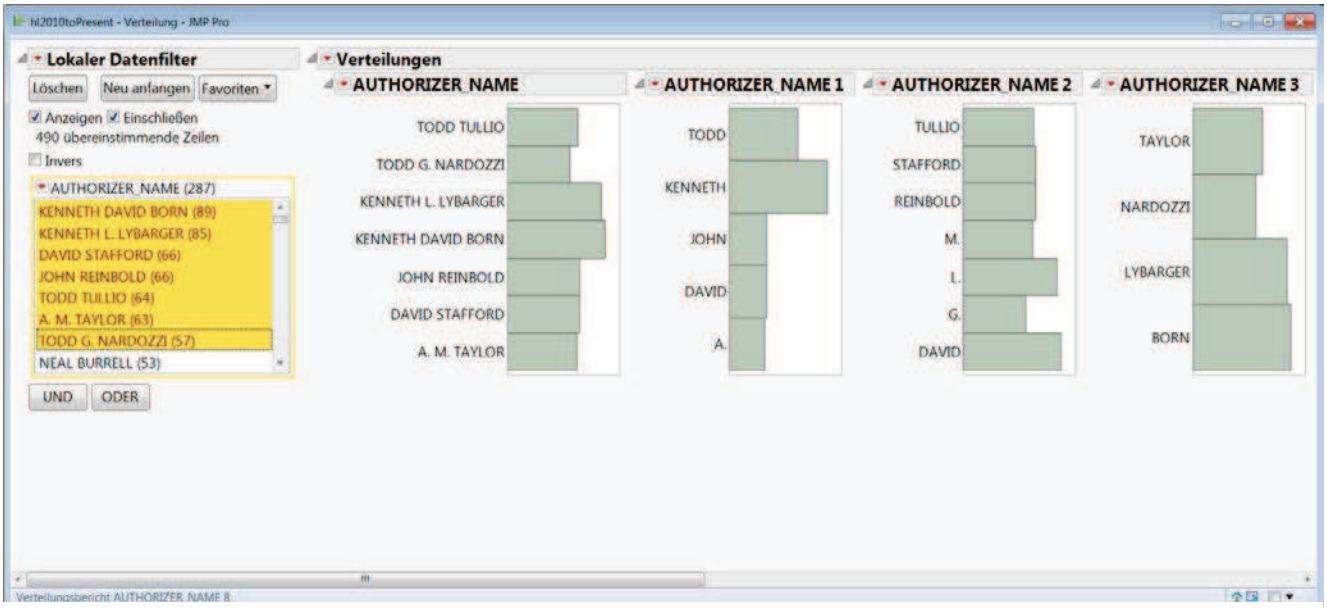



**Abbildung 6:** Trennung nach Worten

Anders ist das Ergebnis der Indikatorfunktion. Hierbei wird für jeden gefundenen Textbestandteil eine Spalte angelegt, die dann nur aus Nullen und Einsen besteht. Für eine Teilmenge des Datensatzes von nur 121 Zeilen zeigt sich, dass Edward A. Fant alleine zu den Textteilen „Edward“ und „Fant“ beiträgt aber nur einen geringen Anteil zu dem Bestandteil „A“ (Abbildung 7).

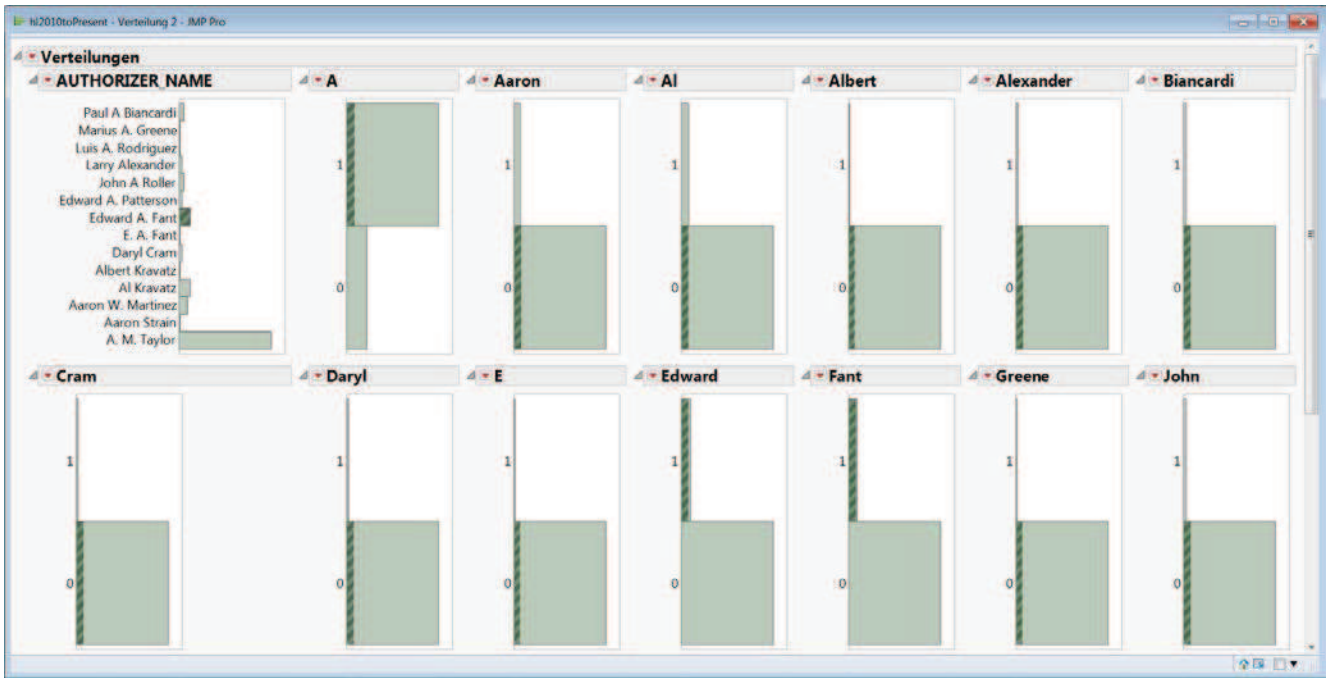



**Abbildung 7:** Indikatorspalten

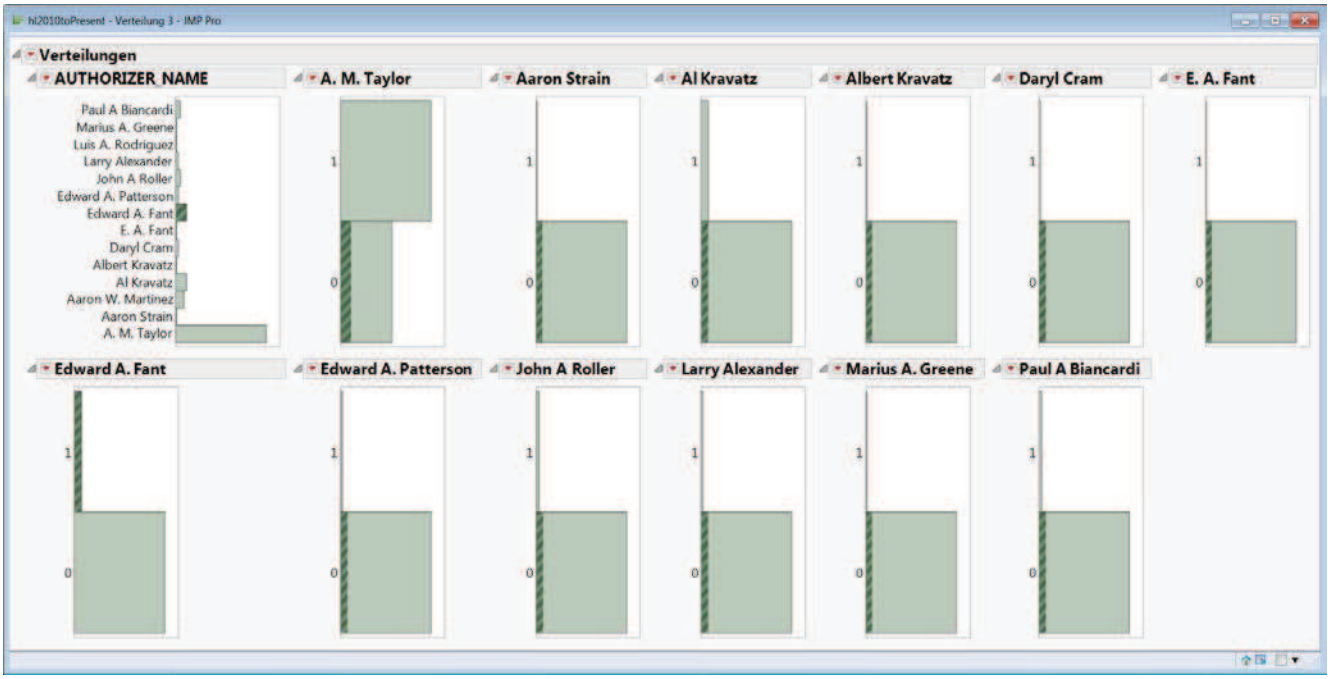



**Abbildung 8:** Indikatorfunktion

## 4.2 Indikatorspalten

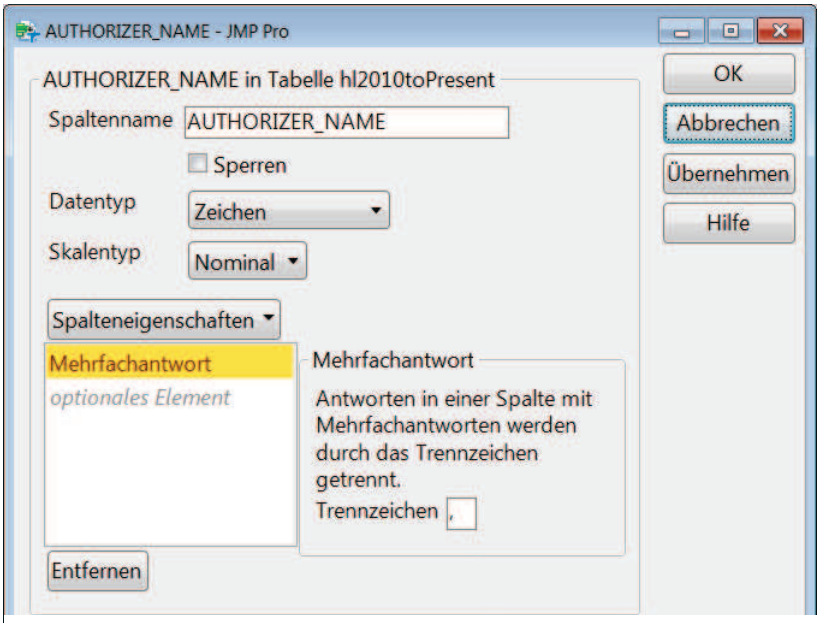



**Abbildung 9:** Spalteneigenschaft Mehrfachantwort

Diese Funktion nimmt den kompletten Text und erzeugt dafür Indikatorspalten (Abbildung 8), Edward A. Fant kommt also nur in „seiner" Indikatorspalte vor. Diese Funktion nutzt allerdings die Spalteneigenschaft „Mehrfachantwort" falls sie angelegt ist (Abbildung 9). Dann sind die Ergebnisse ähnlich wie die der zuvor beschriebenen Indikatorvariante von „Text in Spalten".

## 4.3 Verbinden von Spalten

| AUTHORIZER_NAME | Name Word | Name Indicator | N |
|---|---|---|---|
| A. M. Taylor | A M Taylor | A M Taylor | 77 |
| Aaron Strain | Aaron Strain | Aaron Strain | 1 |
| Aaron W. Martinez | Aaron W Martinez | Aaron Martinez W | 7 |
| Al Kravatz | Al Kravatz | Al Kravatz | 9 |
| Albert Kravatz | Albert Kravatz | Albert Kravatz | 1 |
| Daryl Cram | Daryl Cram | Cram Daryl | 2 |
| E. A. Fant | E A Fant | A E Fant | 1 |
| Edward A. Fant | Edward A Fant | A Edward Fant | 9 |
| Edward A. Patterson | Edward A Patterson | A Edward Patterson | 2 |
| John A Roller | John A Roller | A John Roller | 4 |
| Larry Alexander | Larry Alexander | Alexander Larry | 2 |
| Luis A. Rodriguez | Luis A Rodriguez | A Luis Rodriguez | 1 |
| Marius A. Greene | Marius A Greene | A Greene Marius | 1 |
| Paul A Biancardi | Paul A Biancardi | A Biancardi Paul | 4 |

**Abbildung 10:** Kombinierte Spalten

Inhalte von Zellen können aber nicht nur aufgeteilt sondern auch kombiniert werden. Die zu kombinierenden Spalten werden zunächst in der Tabelle selektiert, dann wird die Funktion aufgerufen. Der zugehörige Dialog erfordert die Angabe eines Spaltennamens für die neue Spalte mit den kombinierten Werten, die Angabe eines Trennzeichens (in Abbildung 9 ist ein Leerzeichen als Trennzeichen eingegeben) und die Information, ob es sich bei den selektierten Spalten um Indikatorspalten handelt. Danach wird die neue Spalte erzeugt. In den vorstehenden Beispielen wurde der Name des Authorizers nach zwei verschiedenen Methoden in seine Bestandteile zerlegt. Einmal erhielt man Spalten mit den Namensbestandteilen, einmal Indikatorspalten. Diese Spalten kann man mit der jeweils passenden Methode zusammenführen. Die Gegenüberstellung des Originalwertes mit den wieder zusammengesetzten Bestandteilen zeigt leichte Unterschiede. Die nach der Wortmethode aufgeteilten Spalten werden wieder so zusammengesetzt, wie der Originaltext

war, bei den nach der Indikatormethode aufgeteilten Spalten werden die Bestandteile in alphabetischer Reihenfolge angegeben (Abbildung 10). Diese Eigenschaft kann man sich zunutze machen, um Textvariablen zu bearbeiten. Wenn der Text z.B. eine Liste von Nennungen bei einem Interview umfasst, lassen sich so die Einträge sortieren, was eine Auszählung eindeutiger Kombinationen erlaubt.

# 5 Modellierungsvorbereitung

Ausreißer, fehlende Werte, Validierung, Identifizieren wichtiger Variablen, sind die Aspekte, die vor jeder Analyse berücksichtigt und gegebenenfalls bearbeitet werden müssen. JMP 12 bietet hierzu entsprechende Funktionen an.

## 5.1 Ausreißer

Ausreißer lassen sich aus vier unterschiedlichen Perspektiven betrachten, die sich zunächst dadurch unterscheiden, dass die Spalten einzeln oder im Zusammenhang betrachtet werden. Je zwei Optionen bieten detaillierte Anpassung.

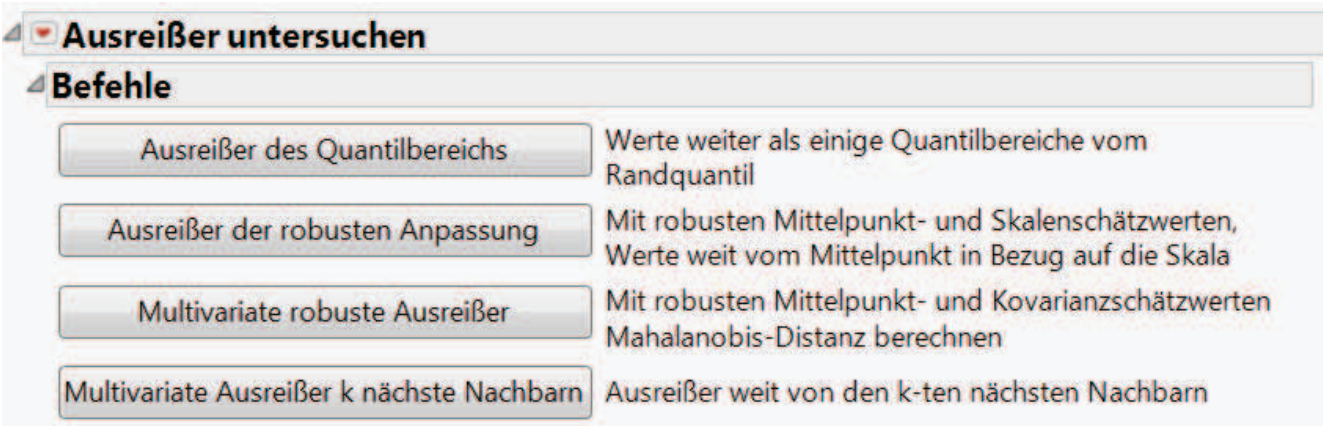
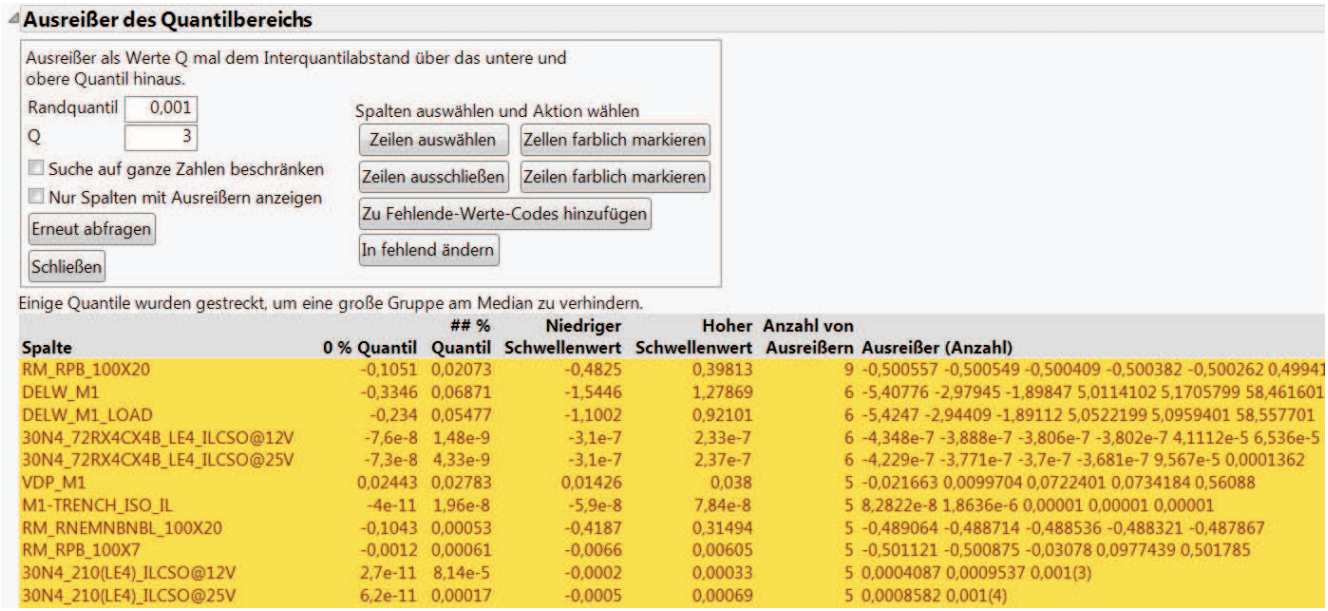



**Abbildung 11:** Ausreißeranalyse

Wird der Quantilsbereich zugrunde gelegt, um Ausreißer zu identifizieren, kann man das Grenzquantil festlegen (in Abbildung 12 das 0,1% Quantil) sowie den Faktor, um den der Abstand einer Beobachtung außerhalb des Quantilsbereichs liegen muss, bevor er als Ausreißer identifiziert wird. Die Ergebnisse dieser Analyse werden aufgelistet.




| Spalte | 0 % Quantil | ## % Quantil | Niedriger Schwellenwert | Hoher Schwellenwert | Anzahl von Ausreißern | Ausreißer (Anzahl) |
|---|---|---|---|---|---|---|
| RM_RPB_100X20 | -0,1051 | 0,02073 | -0,4825 | 0,39813 | 9 | -0,500557 -0,500549 -0,500409 -0,500382 -0,500262 0,49941 |
| DELW_M1 | -0,3346 | 0,06871 | -1,5446 | 1,27869 | 6 | -5,40776 -2,97945 -1,89847 5,0114102 5,1705799 58,461601 |
| DELW_M1_LOAD | -0,234 | 0,05477 | -1,1002 | 0,92101 | 6 | -5,4247 -2,94409 -1,89112 5,0522199 5,0959401 58,557701 |
| 30N4_72RX4CX4B_LE4_ILCSO@12V | -7,6e-8 | 1,48e-9 | -3,1e-7 | 2,33e-7 | 6 | -4,348e-7 -3,888e-7 -3,806e-7 -3,802e-7 4,1112e-5 6,536e-5 |
| 30N4_72RX4CX4B_LE4_ILCSO@25V | -7,3e-8 | 4,33e-9 | -3,1e-7 | 2,37e-7 | 6 | -4,229e-7 -3,771e-7 -3,7e-7 -3,681e-7 9,567e-5 0,0001362 |
| VDP_M1 | 0,02443 | 0,02783 | 0,01426 | 0,038 | 5 | -0,021663 0,0099704 0,0722401 0,0734184 0,56088 |
| M1-TRENCH_ISO_IL | -4e-11 | 1,96e-8 | -5,9e-8 | 7,84e-8 | 5 | 8,2822e-8 1,8636e-6 0,00001 0,00001 0,00001 |
| RM_RNEMNBNBL_100X20 | -0,1043 | 0,00053 | -0,4187 | 0,31494 | 5 | -0,489064 -0,488714 -0,488536 -0,488321 -0,487867 |
| RM_RPB_100X7 | -0,0012 | 0,00061 | -0,0066 | 0,00605 | 5 | -0,501121 -0,500875 -0,03078 0,0977439 0,501785 |
| 30N4_210(LE4)_ILCSO@12V | 2,7e-11 | 8,14e-5 | -0,0002 | 0,00033 | 5 | 0,0004087 0,0009537 0,001(3) |
| 30N4_210(LE4)_ILCSO@25V | 6,2e-11 | 0,00017 | -0,0005 | 0,00069 | 5 | 0,0008582 0,001(4) |

**Abbildung 12:** Ausreißer nach der Quantilsmethode

Wie üblich kann diese Tabelle sortiert werden, z.B. nach der Zahl der pro Spalte gefundenen Ausreißer. Man kann Zeilen der Tabelle markieren, die Zellen mit Ausreißern (für spätere Bearbeitung) farbig markieren und die betroffenen Zeilen (vorerst) von der Analyse ausschließen (Abbildung 13).

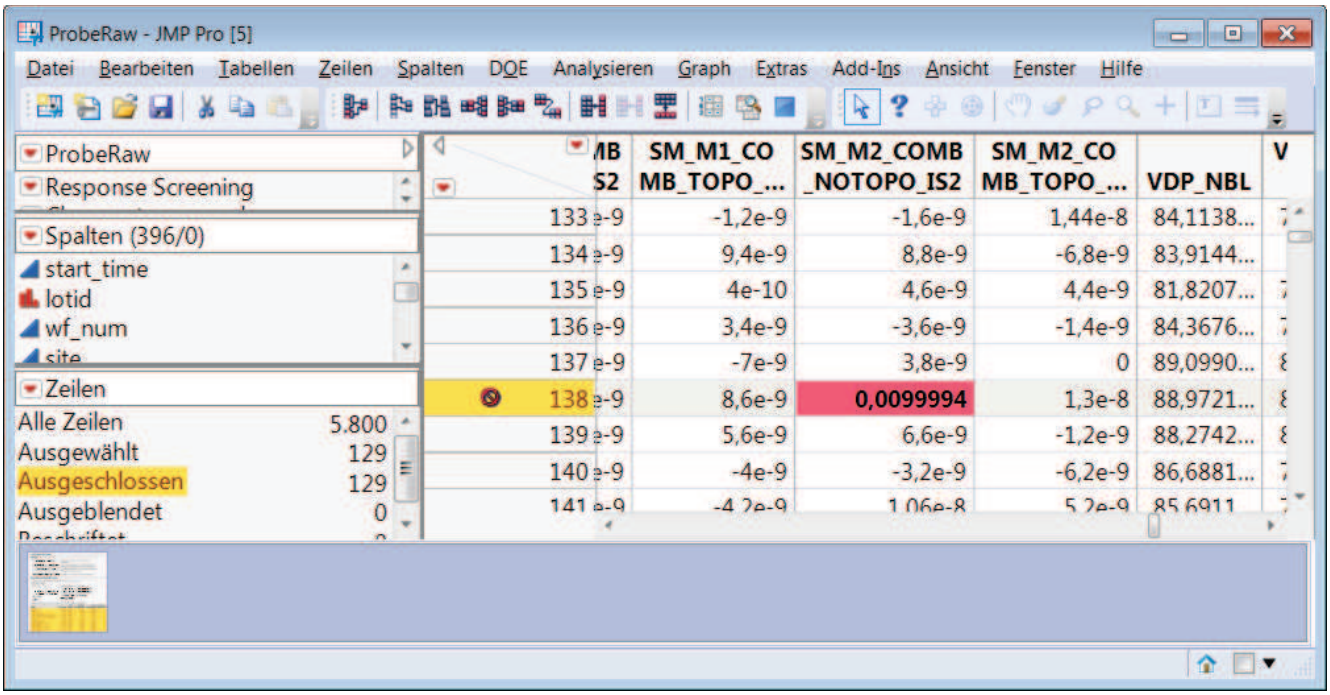



**Abbildung 13:** Markierter und ausgeschlossener Ausreißer

Der Neuner Bericht unterhalb der Tabelle zeigt die Verteilung der jeweils höchsten Neuner Ziffern, üblicherweise ein Code für fehlende Werte. Diese, wie auch die Ausreißer, können in der Tabelle durch fehlende Werte ersetzt werden, die dann spezifisch behandelt werden können.
Um Ausreißer nach der Methode der robusten Anpassung zu erkennen, werden robuste Schätzverfahren (Huber, Cauchy, Quantile) eingesetzt, um Mittelwert und Streuung zu schätzen. Ausreißer liegen dann eine k-fache Streuung vom Mittelwert entfernt.
Die beiden multivariaten Methoden bewerten Beobachtungen nach einem zusammenfassenden Kriterium für mehrere Spalten. Die Methode der robusten multivariaten Ausreißer berechnet die Mahalanobisdistanz jeder Zeile zu dem Mittelwertsvektor. Die Ergebnisse werden als Scatterplot angezeigt. Man kann in diesem Graf Punkte beliebig auswählen und mit den Standardfunktionen von JMP markieren und/oder ausschließen.
Die Berechnung von Abständen nach der Methode der k-nächsten Nachbarn geht davon aus, dass ein Punkt ein Ausreißer ist, wenn er weit entfernt von seinen k nächsten Nachbarn liegt. Der Wert für k wird abgefragt und das Ergebnis der Berechnungen in einem Scatterplot dargestellt. Dieser ist wiederum die „Schaltfläche“ zur Behandlung der Beobachtungen.

## 5.2 Fehlende Werte

Fehlende Werte sind ein häufiges Problem, vor allem wenn man Modelle berechnen will. Bisher fand die Behandlung fehlender Werte in JMP, falls überhaupt, innerhalb einzelner Plattformen statt. JMP Pro hat durchgängig die Option „informativ fehlend“, die fehlende Werte in kategorialen und stetigen Werten unterschiedlich ersetzen. Außerdem gibt es schon länger die Übersicht über das Muster fehlender Werte, aus der heraus man auch Zeilen der Originaltabelle bearbeiten kann. In JMP 12 gibt es nicht nur erweiterte Ansichten über die Struktur der fehlenden Daten, wie Cluster oder eine grafische Darstellung der Verteilung fehlender Werte in der Tabelle, sondern auch die Möglichkeit, fehlende Werte zu ersetzen. Es stehen zwei Methoden zur Wahl.
Die Imputation mittels multivariater Normalverteilung ersetzt fehlende Werte durch eine Regression auf die vorhandenen Werte der Zeile (Abbildung 14). Das Verfahren ist eher geeignet für Tabellen mit einer moderaten Zahl von Spalten und wenn die Annahme einer Normalverteilung nicht grob verletzt ist. Für Tabellen mit vielen Spalten, durchaus tausende, ist die Methode nach der Singulärwertzerlegung eher anwendbar. Durch Berechnung mittels der Pseudoinversen ist der Algorithmus schnell und wird nicht durch Korrelationen beeinträchtigt. Allerdings sinkt die Qualität der Schätzung in dem Maß in dem die Zahl fehlender Werte steigt.

Die Zellen mit ersetzten Werten sind in der Tabelle hellblau unterlegt. Der Vorgang kann wieder rückgängig gemacht werden.


**Abbildung 14:** Analyse fehlender Werte

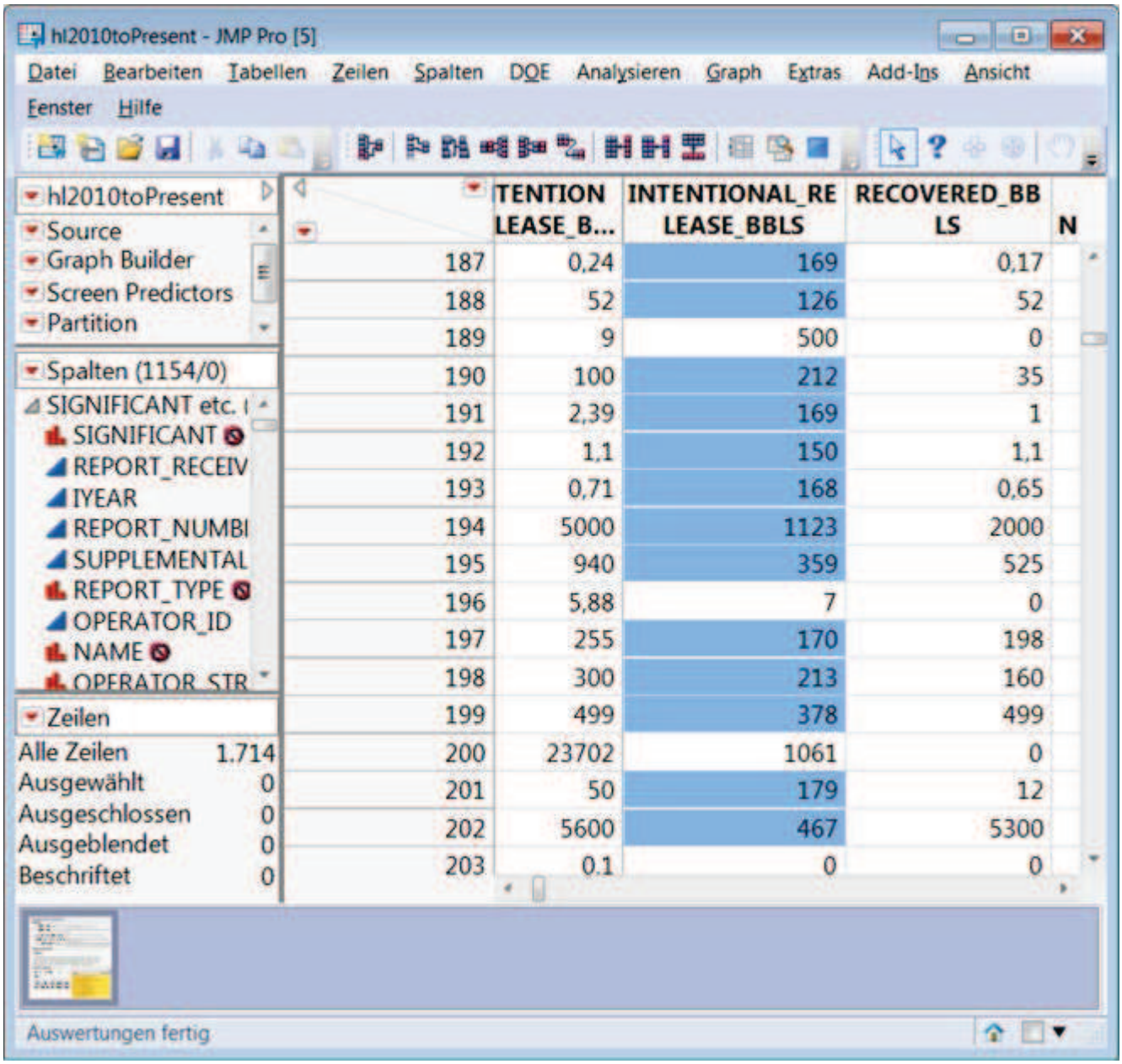



**Abbildung 15:** Ersetzte Zellen

## 6 Fazit

Die spannenden Arbeiten bei der Datenanalyse bleiben natürlich Grafik und Statistik, ebenso wie Neuerungen auf diesen Gebieten auch die spannendsten Fragen bei Vorstellung neuer Programmversionen sind. Dennoch bewähren sich oft die „kleinen Helferlein“ im Alltag besonders. Einsichten in die Mängel der Daten zu bekommen und gleichzeitig auch die Möglichkeit, diese Mängel, wenn nicht zu umgehen, so doch abzuschwächen, erleichtert die Datenvorbereitung deutlich. Hoffentlich bewirken diese Funktionen, dass sich der prozentuale Anteil der Vorarbeiten an dem Gesamtaufwand einer Analyse deutlich reduziert.