



predictum

Data Clean Add-In

minimum requirements:
JMP® 8 (script) /JMP® 9 (add-in)
Copyright © 2011 Predictum Inc.

Features

- Interactive, visible, data cleaning
- Eliminate retest/duplicate data
- Drop mis-test data represented by specific values
- Identify boundary conditions or specify your own
- Pre & post cleaning data characteristics
- Clean stratified data interspersed in the same column

Overall Metrics	
Metric	Value
Number of Columns Cleaned	2
Processing Time Total (s)	13

Column Metrics																
Group	Column	Count	Missing	Total Number Cleaned	Final Count	Percentage Cleaned	Reprocessed Cleaned	Robust Cleaned	Value Cleaned	Boundary Cleaned	Outlier Cleaned	Number of Values	Bounding	Bound Value	Skew	ABS Skew
1	Test 1	90761	0	1967	88794	2.16723	0	1107	66	1	793	6	Lower Bounded	0	.	.
2	Test 3	90761	0	1044	89717	1.15027	0	635	90	102	217	28	Not Bounded	.	.	.

Column Statistics											
Group	Column	Mode	Median	Mid	Mean Post	Sigma Post	CV Post	Variation Post	Range Post	Max Post	Min Post
1	Test 1	0.00005	0.00005	0.00012	5.87e-5	5.28e-5	89.9897	2.79e-9	0.00025	0.00025	0
2	Test 3	13.5533	13.5533	13.535	13.5505	0.11525	0.85053	0.01328	0.9888	14.0294	13.0406



List of data columns

Select columns to be automatically excluded. These column names can be recorded in a global exclusion file as described later in this document.

Specify what should be done with the cleaned data.

Select All button can focus on Continuous columns or any Numeric column regardless of its Modelling Type (Continuous, Ordinal or Nominal).

A variety of cleaning options that are covered later in this document.

Data Clean v4.04 - Demo Mode

Clean numeric data based on options

Select Columns

- Test 1
- Test 3
- Test 4
- Test 5
- Test 6
- Test 7
- Test 8
- Test 9

Columns to Clean (Max 2 in Demo mode)

Select

Select All

Remove

Remove All

- Test 1
- Test 3

Action

OK

Cancel

Recall

Help

Excluded from Select All Options

- Test 2 optional

Options for Select All Button

Continuous Only

Any Numeric

Cleaning Assignment

Set Cleaned to Missing

Set Cleaned to Mean Post Clean

Set Cleaned to a Specific Value

Numeric Value:

Cleaning Options

predictum [Email Feedback](#) www.predictum.com [Download Instructions](#)

Copyright © 2011 by Predictum Inc. [Licensed to Demonstration](#)



Identify a retest/duplicate data column. Eliminate its values when this column has any values

Robust Cleaning: those points beyond that are beyond this region will be cleansed

Specify any particular values that indicate a mis-test

Some data exhibit boundary conditions. Data Clean can auto detect them or you can specify where boundaries are found

Cleaning Options

1) Retest/Duplicated Data Options

Eliminate when this Column has values optional

2) Robust Cleaning

Clean outside X Robust Sigma Value: 8

3) Specific Value Clean List

Clean Values in List

-999
999

4) Boundary/Compliance Clean Options

Do Not use Boundary Conditions
 Auto Detect Boundary Conditions
 Use Specific Boundary Conditions

Use Upper Bound
 Use Lower Bound

5) Outlier Clean Options

Clean Outliers as selected below

Agressive
 Moderate
 Just Some

or, Clean Outliers using a specific Alpha

Diagnostics

Create Diagnostic Columns
 Create Diagnostics Report

Global Options

Create New Table with Cleaned Data
 Create New Table with Purged Data
 Include all coumns in the New Tables
 Clean Locked Columns
 Clean Exlcuded Rows
 Process Formula Columns

not to scale



Choose the level of cleaning to be applied or indicate a specify alpha value for stripping outliers

Choose to create diagnostic columns in the data table and /or diagnostics in the report

Some global options concerning the data table

Cleaning Options

1) Retest/Duplicated Data Options

Eliminate when this Column has values optional

2) Robust Cleaning

Clean outside X Robust Sigma Value: 8

3) Specific Value Clean List

Clean Values in List

-999
999

4) Boundary/Compliance Clean Options

Do Not use Boundary Conditions
 Auto Detect Boundary Conditions
 Use Specific Boundary Conditions

Use Upper Bound
 Use Lower Bound

5) Outlier Clean Options

Clean Outliers as selected below

Agressive
 Moderate
 Just Some

or, Clean Outliers using a specific Alpha

Diagnostics

Create Diagnostic Columns
 Create Diagnostics Report

Global Options

Create New Table with Cleaned Data
 Create New Table with Purged Data
 Include all coumns in the New Tables
 Clean Locked Columns
 Clean Exlcuded Rows
 Process Formula Columns

not to scale



After completing the dialog box, click OK and watch the magic begin!

Data Clean v4.04 - Demo Mode

Clean numeric data based on options

Select Columns

- Test 1
- Test 3
- Test 4
- Test 5
- Test 6
- Test 7
- Test 8
- Test 9

Columns to Clean (Max 2 in Demo mode)

- Test 1
- Test 3

Action

- OK
- Cancel
- Recall
- Help

Excluded from Select All Options

- Test 2 optional

Options for Select All Button

- Continuous Only
- Any Numeric

Cleaning Assignment

- Set Cleaned to Missing
- Set Cleaned to Mean Post Clean
- Set Cleaned to a Specific Value

Numeric Value:

Cleaning Options

predictum [Email Feedback](#) www.predictum.com [Download Instructions](#)

Copyright © 2011 by Predictum Inc. [Licensed to Demonstration](#)



Data Clean Report

▼ Data Clean Report

► Options

▼ Overall Metrics

Metric	Value
Number of Columns Cleaned	2
Processing Time Total (s)	13

▼ Column Metrics

Group	Column	Count	Missing	Total Number Cleaned	Final Count	Percentage Cleaned	Reprocessed Cleaned	Robust Cleaned	Value Cleaned	Boundary Cleaned	Outlier Cleaned	Number of Values	Bounding	Bound Value
1	Test 1	90761	0	1967	88794	2.16723	0	1107	66	1	793	6	Lower Bounded	0
2	Test 3	90761	0	1044	89717	1.15027	0	635	90	102	217	28	Not Bounded	.

▼ Column Statistics

Group	Column	Mode	Median	Mid	Mean Post	Sigma Post	CV Post	Variation Post	Range Post	Max Post	Min Post
1	Test 1	0.00005	0.00005	0.00012	5.87e-5	5.28e-5	89.9897	2.79e-9	0.00025	0.00025	0
2	Test 3	13.5533	13.5533	13.535	13.5505	0.11525	0.85053	0.01328	0.9888	14.0294	13.0406

Reporting the results of the cleaning including the number of observations cleaned from the data with each of the cleaning methods employed



Data Clean Report

- Cleaned Distributions
- Cleaned and Original Distributions**
- Create Limits Table

Metric	Value
Number of Columns Cleaned	2
Processing Time Total (s)	13

Column Metrics

Group	Column	Count	Missing	Total Number Cleaned	Final Count	Percentage Cleaned	Reprocessed Cleaned	Robust Cleaned	Value Cleaned	Boundary Cleaned	Outlier Cleaned	Number of Values	Bounding	Bound Value
1	Test 1	90761	0	1967	88794	2.16723	0	1107	66	1	793	6	Lower Bounded	0
2	Test 3	90761	0	1044	89717	1.15027	0	635	90	102	217	28	Not Bounded	.

Column Statistics

Group	Column	Mode	Median	Mid	Mean Post	Sigma Post	CV Post	Variation						
								Post	Range Post	Max Post	Min Post			
1	Test 1	0.00005	0.00005	0.00012	5.87e-5	5.28e-5	89.9897	2.79e-9	0.00025	0.00025	0			
2	Test 3	13.5533	13.5533	13.535	13.5505	0.11525	0.85053	0.01328	0.9888	14.0294	13.0406			

Select **Cleaned and Original Distribution** from the upper-left red-triangle to launch a report that compares the before and after cleaning



Data Clean Report

Options

Overall Metrics

Metric	Value
Number of Columns Cleaned	2
Processing Time Total (s)	13

Column Metrics

Group	Column	Count	Missing	Total Number Cleaned	Final Count	Percentage Cleaned	Reprocessed Cleaned	Robust Cleaned	Value Cleaned	Boundary Cleaned	Outlier Cleaned	Number of Values	Bounding	Bound Value
1	Test 1	90761	0	1967	88794	2.16723	0	1107	66	1	793	6	Lower Bounded	0
2	Test 3	90761	0	1044	89717	1.15027	0	635	90	102	217	28	Not Bounded	.

Column Statistics

Group	Column	Mode	Median	Mid	Mean Post	Sigma Post	CV Post	Post	Range Post	Max Post	Min Post
1	Test 1	0.0000	0.00005	0.00012	5.87e-5	5.28e-5	89.9897	2.79e-9	0.00025	0.00025	0
2	Test 3	13.5533	13.5533	13.535	13.5505	0.11525	0.85053	0.01328	0.9888	14.0294	13.4046

Distributions - Original and Cleaned

Group 1: Test 1

Column Distributions Original

Column Distributions Cleaned

Group 2: Test 3

Column Distributions Original

Column Distributions Cleaned

Select **Interactive Cleaning** to dynamically see how the various cleaning methods contributed to the exclusion of data

The distribution **before** cleaning

The distribution **after** cleaning



Data Clean Report

Overall Metrics

Metric	Value
Processing Time Total (s)	13

Column Metrics

Group	Column	Count	Missing	Total Number Cleaned	Final Count	Percentage Cleaned	Reprocessed Cleaned	Robust Cleaned	Value Cleaned	Boundary Cleaned	Outlier Cleaned	Number of Values	Bounding	Bound Value
1	Test 1	90761	0	1967	88794	2.16723	0	1107	66	1	793	6	Lower Bounded	0
2	Test 3	90761	0	1044	89717	1.15027	0	635	90	102	217	28	Not Bounded	.

Column Statistics

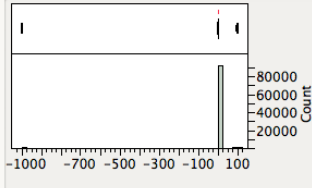
Group	Column	Mode	Median	Mid	Mean Post	Sigma Post	CV Post	Variation Post	Range Post	Max Post	Min Post
1	Test 1	0.00005	0.00005	0.00012	5.87e-5	5.28e-5	89.9897	2.79e-9	0.00025	0.00025	0
2	Test 3	13.5533	13.5533	13.535	13.5505	0.11525	0.85053	0.01328	0.9888	14.0294	13.0406

Distributions - Original and Cleaned

Group 1: Test 1

Column Distributions Original

Test 1



Quantiles	Moments
100.0% maximum: 100.4	Mean: -0.68444
99.5%: 0.0012	Std Dev: 27.008643
97.5%: 0.00025	Std Err Mean: 0.0896576
90.0%: 0.00015	Upper 95% Mean: -0.508717
75.0% quartile: 0.0001	Lower 95% Mean: -0.860355
50.0% median: 0.00005	N: 90761
25.0% quartile: 0	Sum Wgt: 90761
10.0%: 0	Sum: -62120.54
2.5%: 0	Variance: 729.46705
0.5%: 0	Skewness: -36.70256
0.0% minimum: -999	Kurtosis: 1354.5759
	CV: -3946.089
	N Missing: 0

Interactive Cleaning

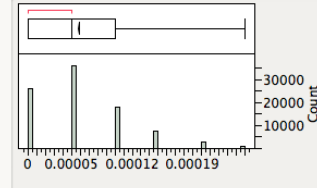
- Retest/Duplicate
- Value List
- Boundary
- Robust
- Outlier

0
66
1107
793

Buttons: Clear Selections, Clean Selected, Add Back Selected, Revert to Original Cleaned, Clear All Cleaning

Column Distributions Cleaned

Test 1

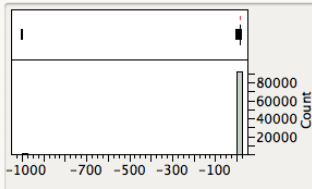


Quantiles	Moments
100.0% maximum: 0.00025	Mean: 5.861
99.5%: 0.00025	Std Dev: 5.271
97.5%: 0.0002	Std Err Mean: 1.777
90.0%: 0.00015	Upper 95% Mean: 0.000
75.0% quartile: 0.0001	Lower 95% Mean: 0.000
50.0% median: 0.00005	N: 8
25.0% quartile: 0	Sum Wgt: 8
10.0%: 0	Sum: 5.208
2.5%: 0	Variance: 2.781
0.5%: 0	Skewness: 0.918
0.0% minimum: 0	Kurtosis: 0.699
	CV: 89.98
	N Missing: 0

Group 2: Test 3

Column Distributions Original

Test 3



Quantiles	Moments
100.0% maximum: 19.9988	Mean: 12.52566
99.5%: 13.8829	Std Dev: 31.874007
97.5%: 13.773	Std Err Mean: 0.105800
90.0%: 13.6998	Upper 95% Mean: 12.733064
75.0% quartile: 13.6265	Lower 95% Mean: 12.318329
50.0% median: 13.5533	N: 90761
25.0% quartile: 13.48	Sum Wgt: 90761
10.0%: 13.4068	Sum: 1136844.7
2.5%: 13.2969	Variance: 1015.9521
0.5%: 12.1616	Skewness: -31.69414
0.0% minimum: -999	Kurtosis: 1002.8609
	CV: 254.46891
	N Missing: 0

Interactive Cleaning

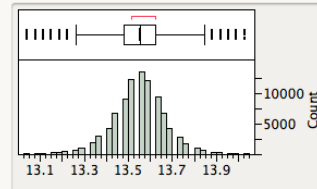
- Retest/Duplicate
- Value List
- Boundary
- Robust
- Outlier

0
90
102
635
217

Buttons: Clear Selections, Clean Selected, Add Back Selected, Revert to Original Cleaned, Clear All Cleaning

Column Distributions Cleaned

Test 3



Quantiles	Moments
100.0% maximum: 14.0294	Mean: 13.59
99.5%: 13.8462	Std Dev: 0.119
97.5%: 13.773	Std Err Mean: 0.000
90.0%: 13.6998	Upper 95% Mean: 13.59
75.0% quartile: 13.6265	Lower 95% Mean: 13.54
50.0% median: 13.5533	N: 8
25.0% quartile: 13.48	Sum Wgt: 8
10.0%: 13.4068	Sum: 1215
2.5%: 13.2969	Variance: 0.013
0.5%: 13.187	Skewness: -0.18
0.0% minimum: 13.0406	Kurtosis: 0.918
	CV: 0.850
	N Missing: 0

Deselect the various cleaning methods and watch the post-cleaning distributions change

The numbers at left indicate how many observations were excluded according to each method



Clean numeric data based on options

Select Columns

- Data 1
- Data 2
- Test Tag 1
- Test Tag 2
- Reprocess optional

Columns to Clean

- Data 1 optional numeric

Action

- OK
- Cancel
- Recall
- Help

Excluded from Select All Options

- optional

Grouped Cleaning

Grouping Columns

- Test Tag 2 optional

Number of Total Groups = 2

Options for Select All Button

- Continuous Only
- Any Numeric

Cleaning Assignment

- Set Cleaned to Missing
- Set Cleaned to Mean Post Clean
- Set Cleaned to a Specific Value

Numeric Value: 0

Cleaning Options

predictum | Email Feedback | www.predictum.com | Download Instructions

Copyright © 2011 by Predictum Inc. | Licensed to Demonstration

Grouped Cleaning is a method to clean one column that has one or more different sources of data

Use it when test data contains before/after in one column or when two or more different process or material conditions are in effect as the data are collected

The highlighted region indicates the number of levels in the grouping column



Data Clean Report

▼ Data Clean Report

► Options

▼ Overall Metrics

Metric	Value
Number of Columns Cleaned	1
Number of Tags	2
Number of Total Groups	2
Processing Time Total (s)	29

▼ Column Metrics

Group	Column	Tag	Count	Missing	Total Number Cleaned	Final Count	Percentage Cleaned	Reprocessed Cleaned	Robust Cleaned	Value Cleaned	Boundary Cleaned	Outlier Cleaned	Number of Values	Bounding	Bound Value
1	Data 1	A	45381	0	783	44598	1.72539	0	668	90	5	20	551	Not Bounded	.
2	Data 1	B	45380	0	781	44599	1.72102	0	666	89	6	20	358	Not Bounded	.

► Column Statistics

 The report expands reporting on each level of the Grouping/Tag column



Testing data can be very large in number of observations and number of parameters tested.

Data Clean makes your life easier by allowing you to specify a universal list of columns that will never be included in analysis.

This list is populated in a simple JMP® file, one as shown in the lower right. It must be named "Data Clean Column Exclusions.jmp"

Place this file in
Mac: user/Library/Application Support/Predictum/

Windows: /Predictum in your home directory

In the dialog in the upper-left, Test 2 is listed from the Exclude from Select All Options because it is listed in the exclusions data table.

The screenshot shows the 'Data Clean v4.04 - Demo Mode' dialog box. It has three main sections: 'Select Columns', 'Columns to Clean (Max 2 in Demo mode)', and 'Action'. In the 'Select Columns' section, a list of columns (Test 1 through Test 9) is shown, with Test 1, 3, 4, 5, 6, 7, 8, and 9 selected. In the 'Columns to Clean' section, Test 1 and Test 3 are listed. In the 'Action' section, there are buttons for OK, Cancel, Recall, and Help. Below the 'Select Columns' section is an 'Excluded from Select All Options' section, which contains 'Test 2 optional'. A 'Grouped Cleaning' section is also visible. In the foreground, the 'Data Clean Column Exclusions.jmp' data table is open, showing a table with 5 rows and 1 column. The third row, containing 'Test 2', is circled in red.

Column 1
1 var1
2 var2
3 Test 2
4 test5
5 test7



Customization?

We're happy to enhance and integrate our Add-Ins to best suit your needs:

- connect directly to your sources of data
- enhanced reporting
- output your reports in PDF or PPT automatically

Check out all our JMP® Add-Ins at
www.predictum.com

