（2022-JA-25MP-16）

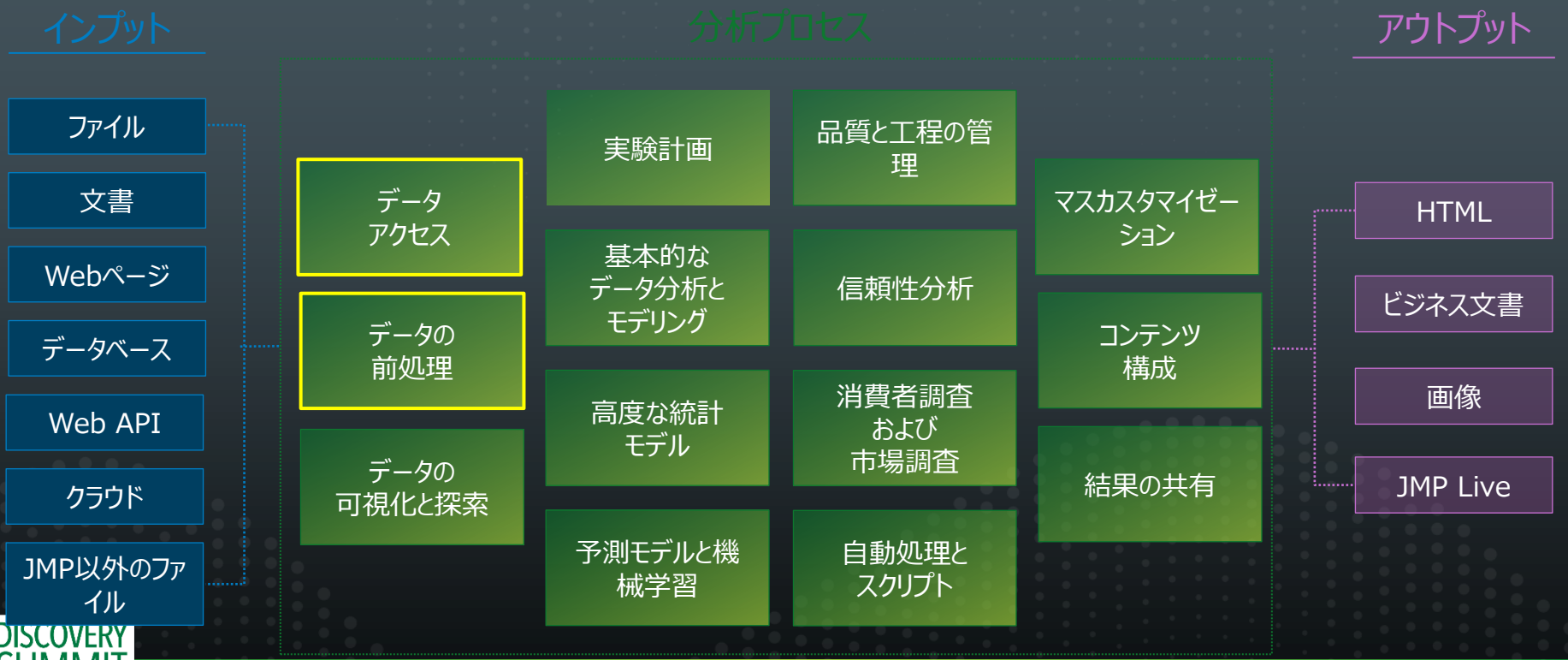
JMP 17で拡張されたデータ加工に関する機能のご紹介

SAS Institute Japan株式会社
原田　大介

JAPAN DISCOVERY SUMMIT
EXPLORING DATA
INSPIRING INNOVATION

jmp

JMPのワークフロー



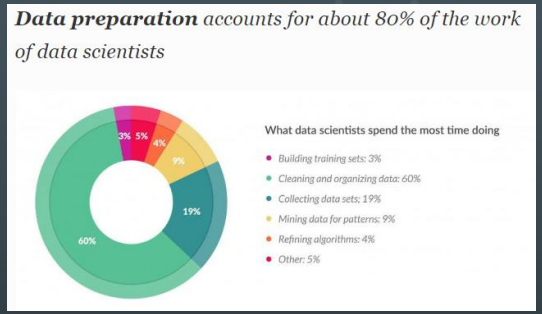
https://www.jmp.com/ja_jp/software/analytic-workflow.html

DISCOVERY SUMMIT
EXPLORING DATA
INSPIRING INNOVATION

jmp

データの前処理には時間がかかる

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says



出展(last accessed 2022/10/19) : Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says (forbes.com)

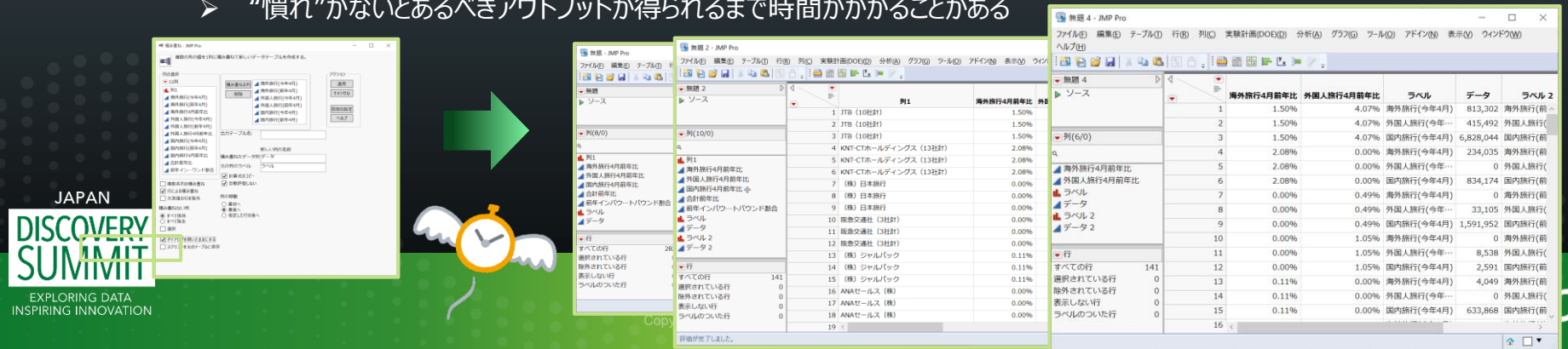
データ分析は前処理が8割、「毒抜き」しないと危険

データの前処理とは、集計や分析に用いる生データを整えて加工すること全般を指します。読者の皆さんの中には、「データ分析は前処理の時間が8割を占める」と聞いたことがある人もいるでしょう。実際、データの前処理は、それだけの時間をかけてでも必要な工程です。

出展(last accessed 2022/10/19) : データ分析は前処理が8割、「毒抜き」しないと危険 | 日経クロステック（xTECH） (nikkei.com)

JMPのデータテーブル加工機能

- JMPのデータテーブルでは種々のデータ加工が施せる
  - 他の表計算ソフトなどで加工/分析はJMPで　といった使い分けはあまり考えなくても良い
- 例えば[テーブル]メニューにある各機能は非常に”使える”
  - 様々なオプションにより柔軟に加工ができる
- 加工をする際にはその加工でやりたいことが出来ることが大前提
  - 様々なオプションの挙動を大まかにでも把握するためには”慣れ”が必要
- JMP16までは「ダイアログを開いたままにする」チェックボックスを使用して、試行錯誤的に加工をする
  - 結果としてデータテーブルが沢山表示されることがある
  - “慣れ”がないとあるべきアウトプットが得られるまで時間がかかることがある

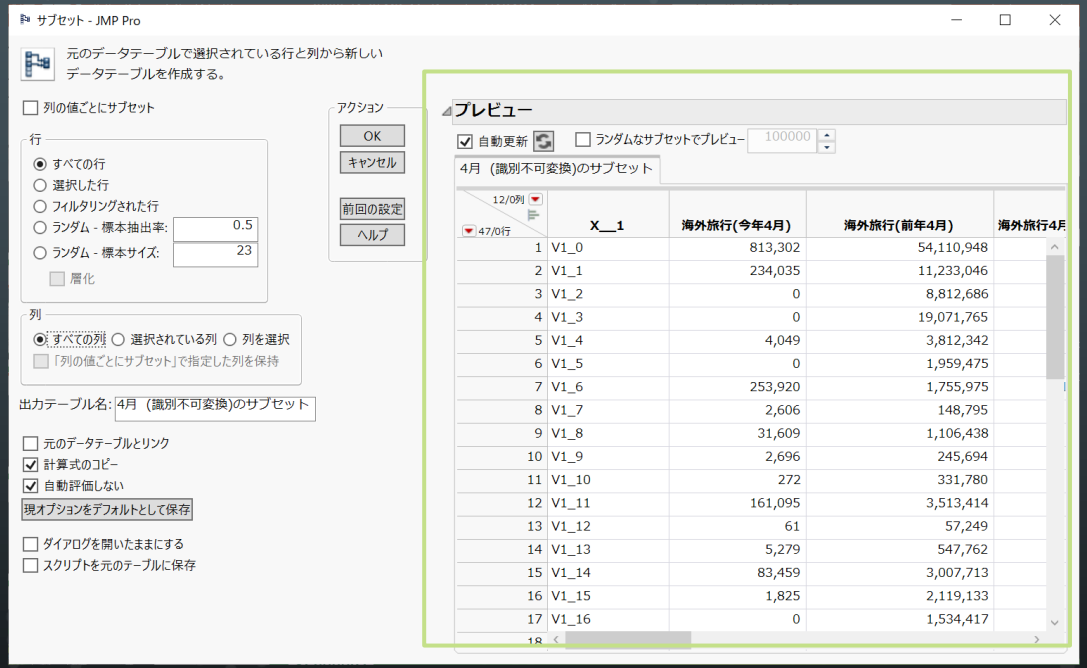


JMPのデータテーブル加工機能

[テーブル]メニュー

要約	さまざまな要約統計量（平均と中央値、標準偏差、最小値と最大値など）が計算して表にまとめる。
サブセット	アクティブなデータテーブルから、すべての行と列、強調表示された行と列、または無作為に選択された行を抽出し、新しいデータテーブルを作成する。
並べ替え	データテーブルを列の値に基づいて昇順または降順に並べ替える。
列の積み重ね	複数の列を1つの新しい列に積み重ね、他の列の値も保持して、データテーブルを構成し直す。
列の分割	1つの列を複数の新しい列に分割して、アクティブなテーブルから新しいデータテーブルを作成する。
転置	データテーブルの行と列を転置させる。
結合(Join)	2つのデータテーブルを1つの新しいテーブルに結合する。
更新	あるデータテーブルの値を別のデータテーブルの値で更新します。列単位で更新ができる。
連結	2つ以上のデータテーブルの行を結合する。

JMP 17では [テーブル]メニューでは加工後の姿がわかる



- 「プレビュー」により加工後のデータテーブルを確認
- 「自動更新」によりオプションを選択した際にプレビューに反映
- 大規模データはランダムなサブセットを自動で作成して速やかにプレビューが更新

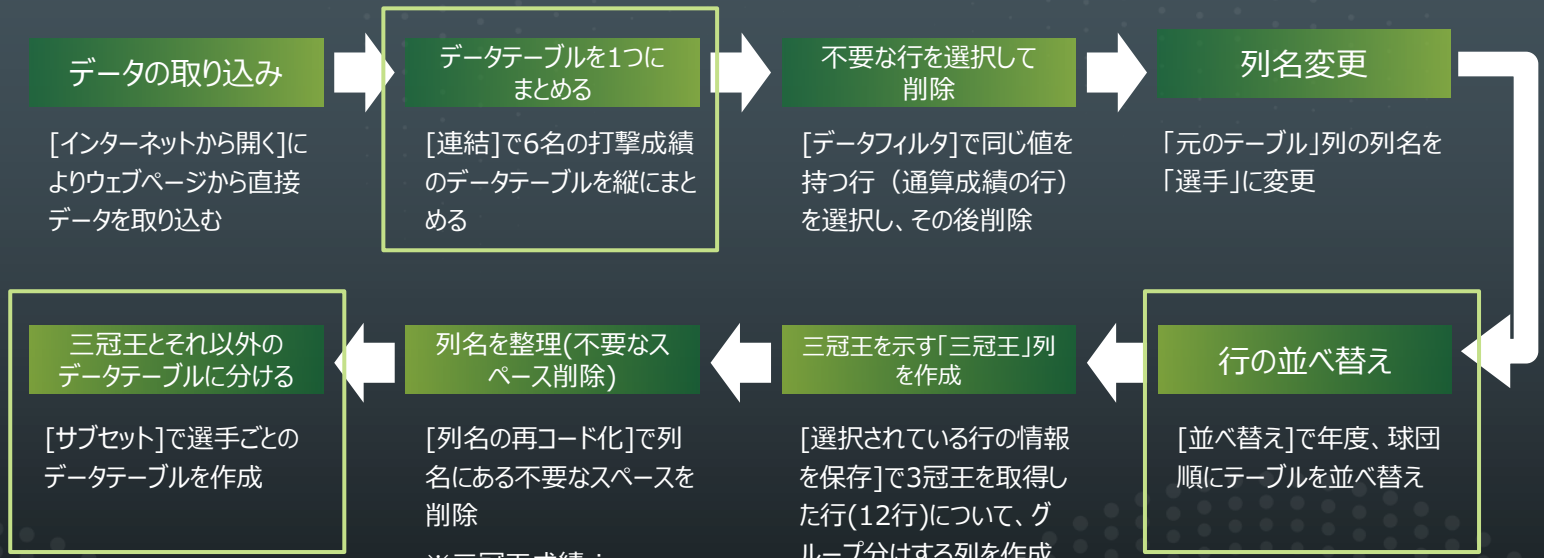
デモ　NPB 歴代三冠王の成績を1つのデータテーブルにまとめる

年度	リーグ	選手名	所属球団
1938年秋	日本野球連盟	中島治康	東京巨人軍
1965年	パ・リーグ	野村克也	南海ホークス
1973年	セ・リーグ	王貞治	読売ジャイアンツ
1974年	セ・リーグ	王貞治	読売ジャイアンツ
1982年	パ・リーグ	落合博満	ロッテオリオンズ
1984年	パ・リーグ	ブーマー・ウェルズ	阪急ブレーブス
1985年	セ・リーグ	ランディ・バース	阪神タイガース
1985年	パ・リーグ	落合博満	ロッテオリオンズ
1986年	セ・リーグ	ランディ・バース	阪神タイガース
1986年	パ・リーグ	落合博満	ロッテオリオンズ
2004年	パ・リーグ	松中信彦	福岡ダイエーホークス
2022年	セ・リーグ	村上宗隆	東京ヤクルトスワローズ

出展※：三冠 (野球) - Wikipedia

- 歴代3冠王8名のデータはWikipediaやNPBのウェブサイトを見ると打率、打点、本塁打についてはまとめてある
- その他の打撃成績も一つの表にまとめたい
  - **<u>各選手の現役時年度別打撃成績をまとめる</u>**
  - ***<u>Wikipedia</u>*<u>に掲載の内容をデータソースとする</u>**

デモ　加工の流れ



(大規模)データからランダムに行を選択して取り込む機能

何十万何百万行もある大きなデータテーブルが手元にあるとき…

- JMPはファイルをパソコンのメモリに展開するため開くとメモリを消費する
- メモリを消費するだけでなく、ファイルを開くことがそもそも大変(時間がかかる)

例）3,000,000行×1000列のデータテーブルを開いたところ**約70秒**要する

- どんなデータだったか中身を少し見てみたい
- ランダムに抽出して分析したい

JMP 17の新機能（JSL限定）

- 最初のn行を指定して開く
- 最後のn行を指定して開く
- ランダムに行を選択して開く

JAPAN DISCOVERY SUMMIT
EXPLORING DATA INSPIRING INNOVATION

(大規模)データからランダムに行を選択して取り込む機能

- ファイルのすべてを開く

```
dt = Open( filepath );
```

- データテーブル最初のn行を開く

```
dt = Open( filepath , First( n ) );
```

- データテーブル最後のn行を開く

```
dt = Open( filepath , Last( n ) );
```

- データテーブルの全行から割合p(pは0～1)でランダムに開く

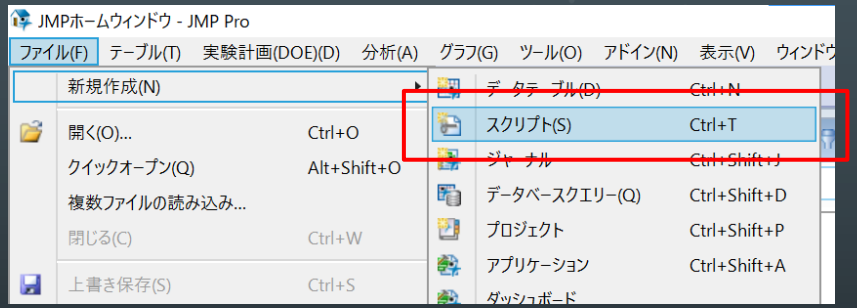
```
dt = Open( filepath , Random( p ) );
```

※filepathは”C:¥data.jmp”のように2重引用符で囲ったものを記述

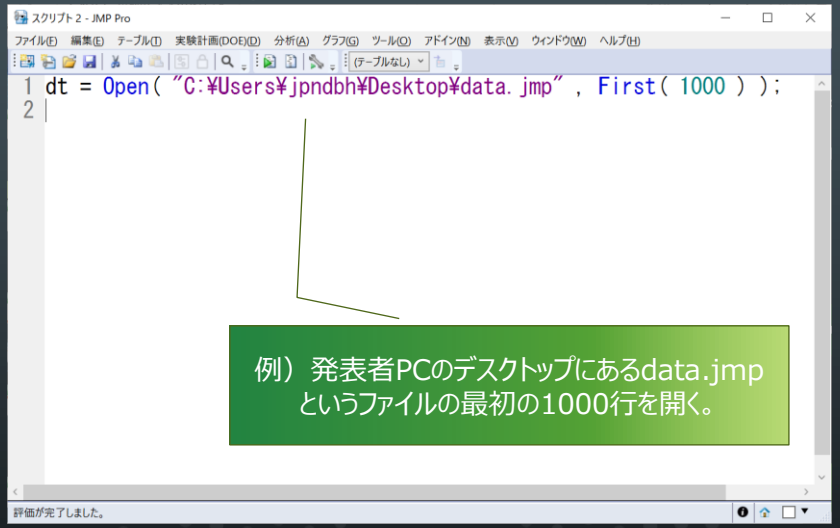
現時点ではJSL(JMPスクリプト言語)限定の
機能です。GUIではこの機能は使用できません。

デモ　デスクトップにある3M行×1K列のファイルを開いてみる

①[ファイル]>[新規作成]>[スクリプト]からスクリプトウィンドウを開く



②前スライドに記載のスクリプトを記述する



③[編集]>[スクリプトの実行] を選択(もしくはCtrl+rをクリック)

現時点ではJSL(JMPスクリプト言語)限定の
機能です。GUIではこの機能は使用できません。

JAPAN
DISCOVERY SUMMIT
EXPLORING DATA
INSPIRING INNOVATION

ご清聴ありがとうございました

jmp

参考

JMPの[テーブル]メニューでできるデータ加工の例

要約：さまざまな要約統計量（平均と中央値、標準偏差、最小値と最大値など）が計算して表にまとめる

Group	X1	X2
A	5	10
A	5	6
A	4	5
A	2	5
A	5	10
B	7	10
B	2	10
B	1	7
B	4	5
B	6	7

- Group毎のX1とX2の平均値や標準偏差

Group	行数	平均(X1)	平均(X2)	標準偏差(X1)	標準偏差(X2)
A	5	4.2	7.2	1.303840481	2.588435821
B	5	4	7.8	2.549509757	2.167948339

JMPの[テーブル]メニューでできるデータ加工の例

サブセット：アクティブなデータテーブルから、すべての行と列、強調表示された行と列、または無作為に選択された行を抽出し、新しいデータテーブルを作成できます。

X1	X2	X3	X4
43	2	22	6
42	10	2	4
35	8	29	8
57	3	23	4
16	10	30	8
78	3	23	7
87	7	11	5
56	5	27	3
51	6	1	5
53	5	6	1

X2	X3
10	2
8	29
3	23
6	1

- 列X2、X3の2，3，4，9行目を抽出

JMPの[テーブル]メニューでできるデータ加工の例

列の積み重ね：複数の列を1つの新しい列に積み重ね、他の列の値も保持して、データテーブルを構成し直します。

X1	X2	X3	X4
43	2	22	6
42	10	2	4
35	8	29	8
57	3	23	4
16	10	30	8

＋

22	6
2	4
29	8
23	4
30	8

- X1＆X3、X2&X4の隣接2系列を積み重ね

JMPの[テーブル]メニューでできるデータ加工の例

列の分割：1つの列を複数の新しい列に分割して、アクティブなテーブルから新しいデータテーブルを作成できます。

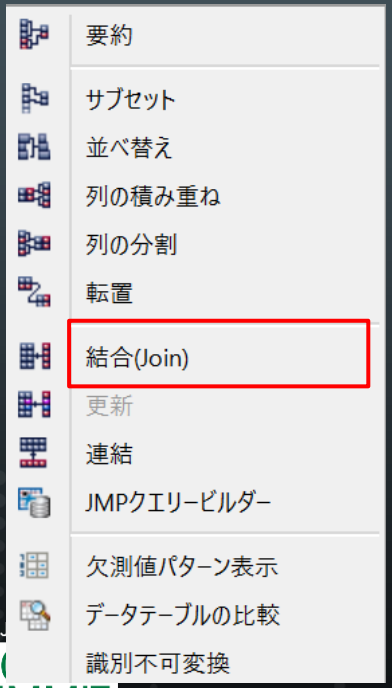
Group	X1	X2
A	5	10
A	5	6
A	4	5
A	2	5
A	5	10
B	7	10
B	2	10
B	1	7
B	4	5
B	6	7

X1 A	X1 B	X2 A	X2 B
5	7	10	10
5	2	6	10
4	1	5	7
2	4	5	5
5	6	10	7

- X1とX2をGroup列の値を基準に分割

JMPの[テーブル]メニューでできるデータ加工の例

[テーブル]メニュー



結合(Join)：2つのデータテーブルを1つの新しいテーブルに結合できます。

ID	X1	X2
1	2	4
2	8	7
3	4	3
4	8	6

＋

ID	X3	X4
3	4	8
4	2	2
2	5	7
1	5	5

↓

ID	X1	X2	X3	X4
1	2	4	5	5
2	8	7	5	7
3	4	3	4	8
4	8	6	2	2

- 2つのデータテーブルのID列の値で対応させて結合

JMPの[テーブル]メニューでできるデータ加工の例

[テーブル]メニュー

連結：2つ以上のデータテーブルの行を結合します。

ID	X1	X2
1	2	4
2	8	7
3	4	3
4	8	6

＋

ID	X1	X2
3	4	8
4	2	2
2	5	7
1	5	5

⇒

ID	X1	X2
1	2	4
2	8	7
3	4	3
4	8	6
3	4	8
4	2	2
2	5	7
1	5	5

- 2つの列数や列名が同じデータテーブルの行を結合

デモで使用したデータ加工機能のまとめ

インターネットから開く	[ファイル]>[インターネットから開く]>[Webページ]。インターネットやリモートコンピュータからデータやイメージを読み込み、それをデータテーブル、Webページ、またはテキストとして保存できます。
データフィルタ	[行]>[データフィルタ]。データの様々なサブセット（一部分）を対話的に選択し、これらのサブセットをプロット上で非表示にしたり、分析から除外したりできます。同様の機能に[ローカルデータフィルタ]があります。こちらはレポートウィンドウで使用できます。
置換	[編集]>[検索]>[検索]。セルの値を検索および置換できます。列名を検索することなども出可能です。
列名変更	表頭で対象列名を選択した状態でタイプします。[列]>[列情報]からの変更も可能です。
再コード化	[列]>[再コード化]。列内のすべての値を一度に変更することができます。グループ単位での変更もできるため、複数の水準を容易にまとめることが出来ます。
並べ替え	[テーブル]>[並べ替え]。列の値に基づいて昇順または降順に並べ替えることができます。
選択されている行の情報を保存	[行]>[行の選択]>[選択されている行の情報を保存]。現在の行の選択を新しい列に保存することができます。例）選択している行=1、選択していない行=2という値を取る列を作成できます。
列名の再コード化	[列]>[列名]>[列名の再コード化]。列見出しの再コード化を行うことが出来ます。

※[テーブル]メニューの機能は前頁までを参照