

人工智慧在生活的應用 常見的資料前處理技術



國立宜蘭大學資訊工程系

吳政瑋 助理教授

wucw@niu.edu.tw



從Excel與CSV檔案中匯出資料



學習目標

- 瞭解什麼是 CSV 格式
- 使用 File 元件
- 使用 Data Table 元件



認識CSV與TSV格式

- ✓ CSV (Comma-Separated Values)
 - 用半形逗號(,)區隔資料欄位
 - 文字內容不可以包含逗號
 - 文字內容必需用逗號時，可以使用引號(“,”)包裹
 - 一般情況中，第一行經常作為標題列(Column Name)

CSV範例：

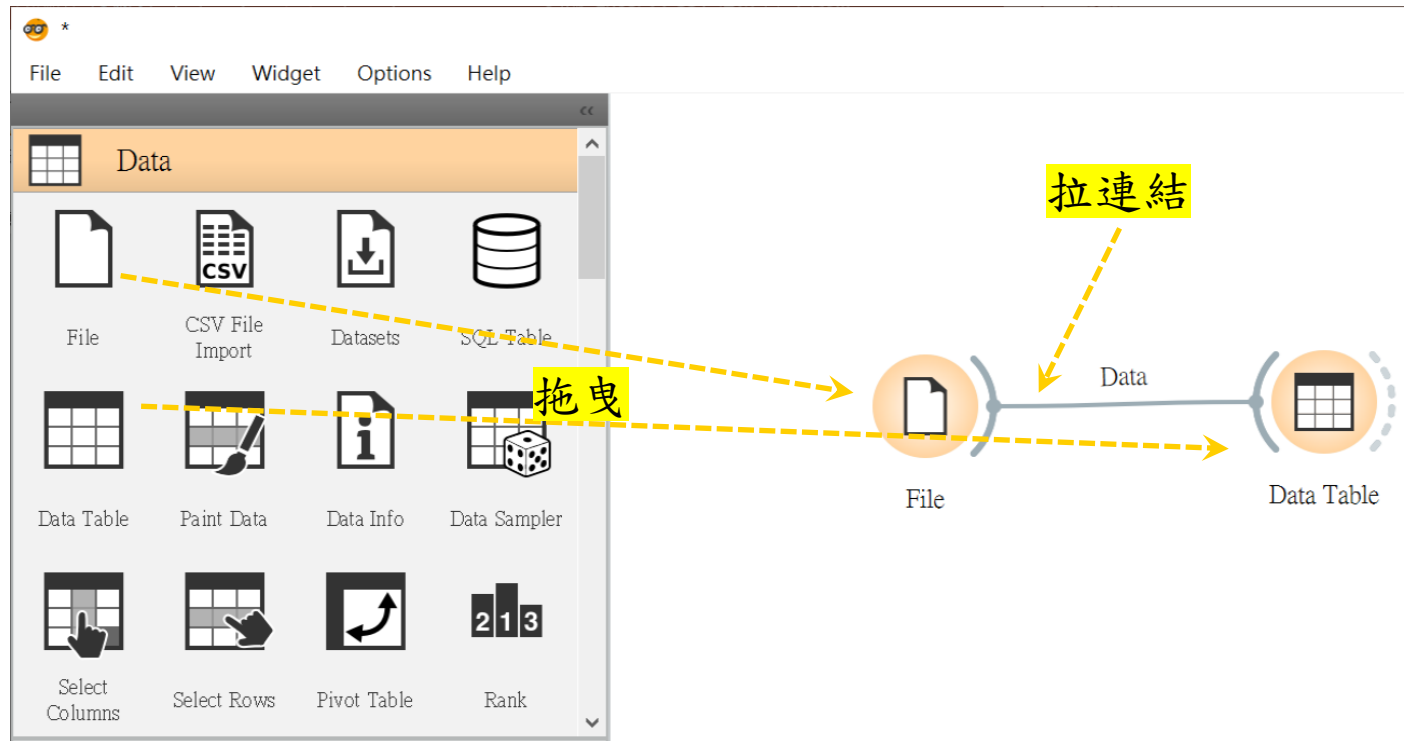
Title, author, ISBN13
Animal Farm, George Orwell, 9837261738
Brave New World, Aldous Huxley, 9835122775
Jurassic Park, Daniel Wu”,”Victor Wu, 9853657268
.
.
.



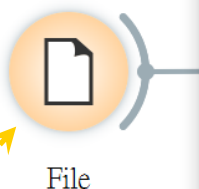
實際操作

- 使用 File 及 Data Table 元件建立資料流程
 - 拖曳 File 及 Data Table 到工作區
 - 從 File(輸出) 拉流程連結線到 Data Table(輸入)
 - 操作 File 元件
 - 操作 Data Table 元件

拖曳 File 及 Data Table 到工作區





操作 File 元件



點擊兩下開啟
File 元件編輯視窗

File

☒ File: iris.tab  

☐ URL:

Info

Iris flower dataset
Classical dataset with 150 instances of Iris setosa, Iris virginica and Iris versicolor.

150 instance(s)
4 feature(s) (no missing values)
Classification; categorical class with 3 values (no missing values)
0 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	sepal length	N numeric	feature	
2	sepal width	N numeric	feature	
3	petal length	N numeric	feature	
4	petal width	N numeric	feature	
5	iris	C categorical	target	Iris-setosa, Iris-versicolor, Iris-...

點擊後可選取要
引入的檔案，常用
的格式有：

*.CSV
*.TSV
*.XLSX

或輸入URL鍊結點
引入檔案

點擊兩下可更改欄位(column)的
數據型態

操作 Data Table 元件

點擊一下欄位名稱，全部資料會以升冪▲(Ascending)方式排列

再點擊一下欄位名稱，全部資料會以降冪▼(Descending)方式排列



File

Data



Data Table

點擊兩下開啟
Data Table 元件
編輯視窗

Data Table

Variables

- ☒ Show variable labels (if present)
- ☐ Visualize numeric values
- ☒ Color by instance classes

Selection

- ☒ Select full rows

點擊一下可回復到
資料初始排列順序

Restore Original Order

- ☒ Send Automatically

? | 150

	iris	sepal length	sepal width	petal length	petal wic
14	Iris-setosa	4.3	3.0	1.1	
43	Iris-setosa	4.4	3.2	1.3	
39	Iris-setosa	4.4	3.0	1.3	
9	Iris-setosa	4.4	2.9	1.4	
42	Iris-setosa	4.5	2.3	1.3	
48	Iris-setosa	4.6	3.2	1.4	
23	Iris-setosa	4.6	3.6	1.0	
7	Iris-setosa	4.6	3.4	1.4	
4	Iris-setosa	4.6	3.1	1.5	
30	Iris-setosa	4.7	3.2	1.6	
3	Iris-setosa	4.7	3.2	1.3	
46	Iris-setosa	4.8	3.0	1.4	
31	Iris-setosa	4.8	3.1	1.6	
25	Iris-setosa	4.8	3.4	1.9	
13	Iris-setosa	4.8	3.0	1.4	
12	Iris-setosa	4.8	3.4	1.6	
107	Iris-virginica	4.9	2.5	4.5	
58	Iris-versicolor	4.9	2.4	3.3	

* 所有 Orange 3 的
操作都是基於表
格

* Data Table 元件主
要的作用是看表
格

操作 Data Table 元件


The screenshot shows the Data Table component interface. On the left, a diagram illustrates the data flow: a 'File' icon (document) feeds into a 'Data' stream, which then feeds into the 'Data Table' component (table icon). A yellow box highlights the 'Visualize numeric values' checkbox in the 'Variables' section, with a red dashed border around it. A yellow arrow points from this checkbox to the 'sepal length' column in the data table. Below the yellow box, text reads: '勾選 Visualize numeric Values 可觀看整體數值的趨勢分布'.

Data Table Settings:

- Variables:**
 - ☒ Show variable labels (if present)
 - ☒ Visualize numeric values
 - ☒ Color by instance classes
- Selection:**
 - ☒ Select full rows
- Buttons:** Restore Original Order, Send Automatically (checked)

Data Table Content:

	iris	sepal length	sepal width	petal length	petal v
14	Iris-setosa	4.3	3.0	1.1	0.1
43	Iris-setosa	4.4	3.2	1.3	0.2
39	Iris-setosa	4.4	3.0	1.3	0.2
9	Iris-setosa	4.4	2.9	1.4	0.2
42	Iris-setosa	4.5	2.3	1.3	0.3
23	Iris-setosa	4.6	3.6	1.0	0.2
7	Iris-setosa	4.6	3.4	1.4	0.3
48	Iris-setosa	4.6	3.2	1.4	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
30	Iris-setosa	4.7	3.2	1.6	0.2
12	Iris-setosa	4.8	3.4	1.6	0.2
25	Iris-setosa	4.8	3.4	1.9	0.2
31	Iris-setosa	4.8	3.1	1.6	0.2
13	Iris-setosa	4.8	3.0	1.4	0.1
46	Iris-setosa	4.8	3.0	1.4	0.3
10	Iris-setosa	4.9	3.1	1.5	0.1
35	Iris-setosa	4.9	3.1	1.5	0.1



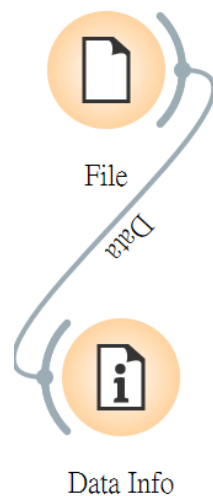
資料表操作： 摘要、選擇行、選擇列



學習目標

- 使用 Data Info 元件
- 使用 Select Rows 元件
- 使用 Select Columns 元件

Data Info 元件說明



Data Info

Data Set Name: iris **資料名稱**

Data Set Size: Rows: 150, Columns: 5 **行列數量**

Features: Categorical: -, Numeric: 4 **類別數量**

Targets: Categorical outcome with 3 values **目標資訊**

Meta Attributes: None **Meta屬及數量**

Location: Data is stored in memory **資料存在位置**

Data Attributes: Name: Iris flower dataset, Description: Classical dataset with 150..., Author: Edgar Anderson, Ronald Fisher, Year: 1936, Reference: R. A. Fisher (1936). "The..." **資料屬性**

? | 150

Data Info的用途
是呈現資料狀態

Discrete:離散型態
Numeric:數值型態
Textual:字串型態

Select Rows 元件說明

Select Rows

Conditions

篩選欄位選擇：
選擇加入篩選條件式的欄位名稱

公式式選擇：
選擇篩選條件式的公式

數值輸入：
輸入套入公式的篩選數值

Chinese equals

equals
is not
is below
is at most
is greater than
is at least
is between
is outside
is defined

equals 等於 =
is not 不等於 ≠
is below 小於 <
is at most 小於等於 ≤
is great than 大於 >
is at least 大於等於 ≥
is between 介於兩數值間
is outside 除了兩數值以外
is defined 被定義的...

Add Condition Add All Variables Remove All

☐ Remove unused features
☐ Remove unused classes

☒ Send Automatically

? | 6 6

Select Rows 元件說明

Select Rows

Conditions

<input checked="" type="checkbox"/> Chinese	is greater than	80	
<input checked="" type="checkbox"/> Math	is between	50	and 70
<input checked="" type="checkbox"/> English	is greater than	80	

☐ Remove unused features

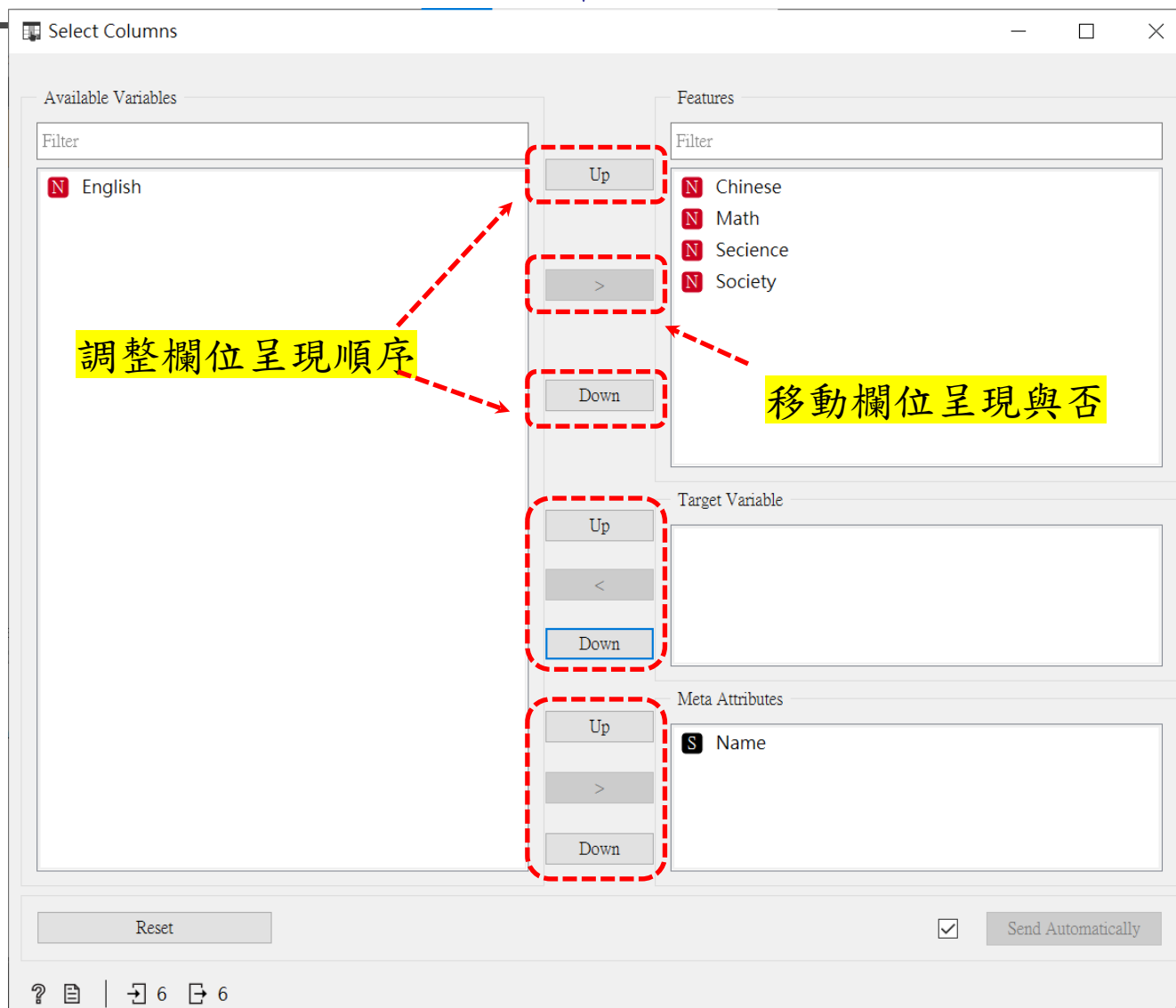
☐ Remove unused classes

☒ **新增篩選條件式**

☒ **全部欄位加入篩選條件**

? | 6 | 1

Select Columns 元件說明



Select Columns:

用於控制每個欄位
是否是否呈現？
呈現的位置？

元件資料Summary與Report

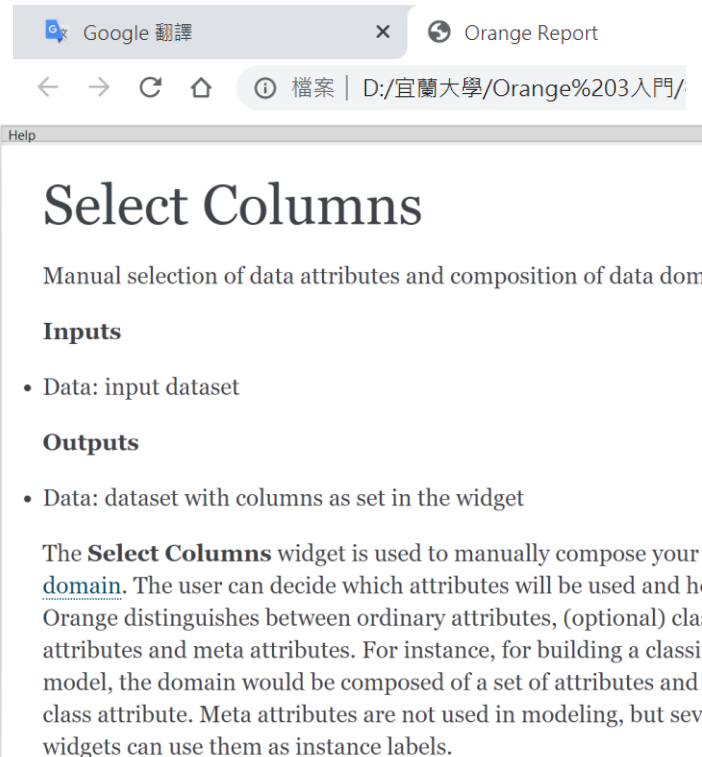
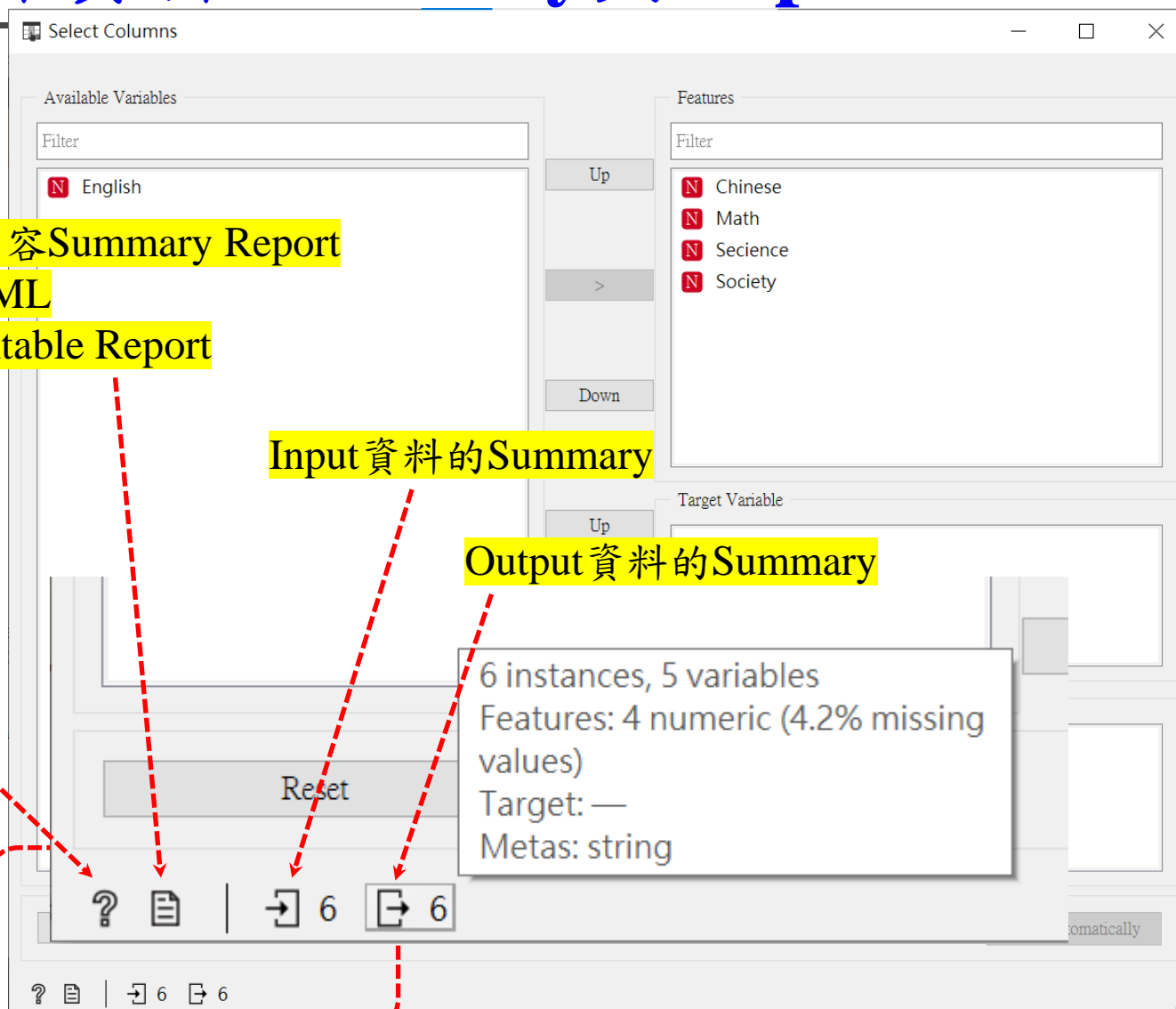
元件內容Summary Report

- HTML
- Printable Report

Input資料的Summary

Output資料的Summary

顯示本元件詳細
說明於工作區



元件詳細說明

Summary Report - HTML



實際操作

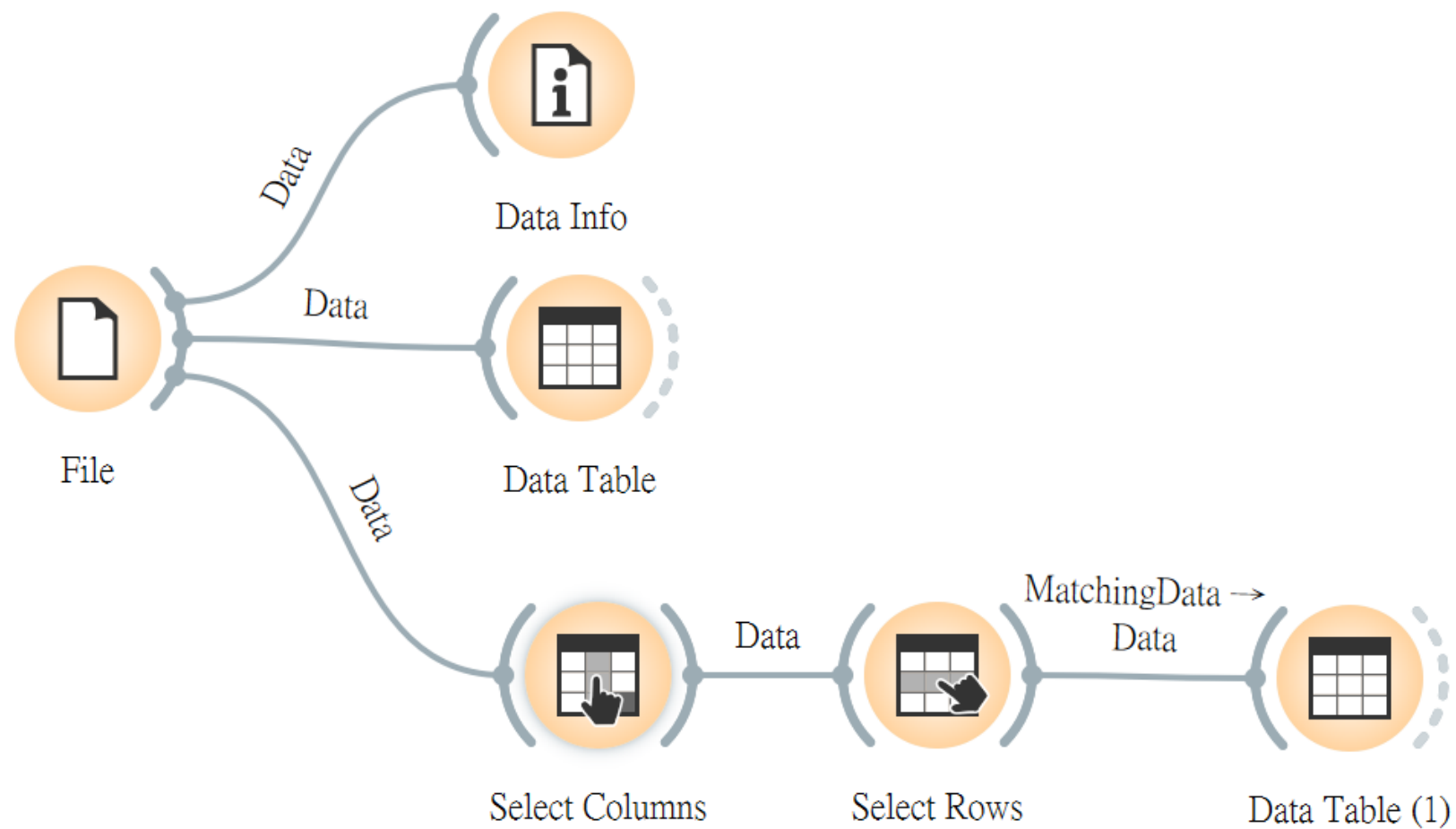
✓ 範例說明

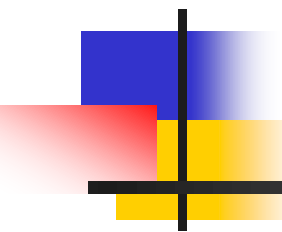
- 小學各科成績檔 Score.CSV
- 包含Chinese, Math, English, Science, Society 五科成績

✓ 使用 File 及 Data Table 元件建立資料流程

1. 建立資料流程圖
2. File元件導入Score.csv
3. 情境練習:篩選出語言成績優秀，數學成績不佳，但數學成績可加強補救，參加數學補救教學的學生。

建立資料流程圖





資料表操作： 離散化、連續化

學習目標：

- 理解連續化、離散化的目的
 - 使用 Continuize 元件
 - 使用 Discretize 元件



學習目標

- 理解連續化、離散化的目的
- 使用 Continuize 元件
- 使用 Discretize 元件

Continuze 連續化

連續化(Continuze)是為了將資料傳送到某些特定的模型當中

ID	Level
01	high
02	high
03	low
04	medium

原始資料

轉換



ID	high	medium	low
01	1	0	0
02	10	0	0
03	0	0	1
04	0	1	0

連續化資料資料



Discretize 離散化

- 離散化(Discretize)就是建立分組
- 常用的兩種分組方式
 - 依照分佈百分比分組 (例如分四組: PP25 - PP50 - PP75)
 - 依照固定間距分組：設定分組數後，由Orange 3界定範圍

Discretize – Equal-frequency (分佈百分比)

Discretize

Default Discretization

☒ Equal-frequency discretization

Num. of intervals: 3

☐ Equal-width discretization

Individual Attribute Settings

☒ Chinese: 79.50, 86.00

☒ Math: 61.50, 77.00

☒ English: 73.00, 87.50

☒ Science: 68.50, 92.50


☒ Society: 83.50, 87.50

	Name	Chinese	Math	English	Science	Society
1	Anna	≥ 86	61.5 - 77	≥ 87.5	68.5 - 92.5	83.5 - 87.5
2	Benson	< 79.5	≥ 77	73 - 87.5	68.5 - 92.5	< 83.5
3	Christine	79.5 - 86	< 61.5	< 73	< 68.5	?
4	David	79.5 - 86	≥ 77	73 - 87.5	≥ 92.5	≥ 87.5
5	Fanny	< 79.5	< 61.5	< 73	< 68.5	< 83.5
6	Eason	≥ 86	≥ 77	≥ 87.5	≥ 92.5	≥ 87.5

結果：Data Table 元件

設定：Discretize 元件

Discretize – Equal-width (固定間距)

 Discretize






Default Discretization

☐ Equal-frequency discretization

☒ Equal-width discretization

Num. of intervals:

Individual Attribute Settings

-  Chinese: 69.33, 83.67
-  Math: 62.00, 79.00
-  English: 65.33, 78.67
-  Science: 64.00, 80.00
-  Society: 69.00, 83.00

	Name	Chinese	Math	English	Science	Society
1	Anna	≥ 83.67	62 - 79	≥ 78.67	64 - 80	≥ 83
2	Benson	69.33 - 83.67	≥ 79	≥ 78.67	≥ 80	69 - 83
3	Christine	69.33 - 83.67	< 62	65.33 - 78.67	< 64	?
4	David	≥ 83.67	≥ 79	≥ 78.67	≥ 80	≥ 83
5	Fanny	< 69.33	< 62	< 65.33	< 64	< 69
6	Eason	≥ 83.67	≥ 79	≥ 78.67	≥ 80	≥ 83

結果：Data Table元件

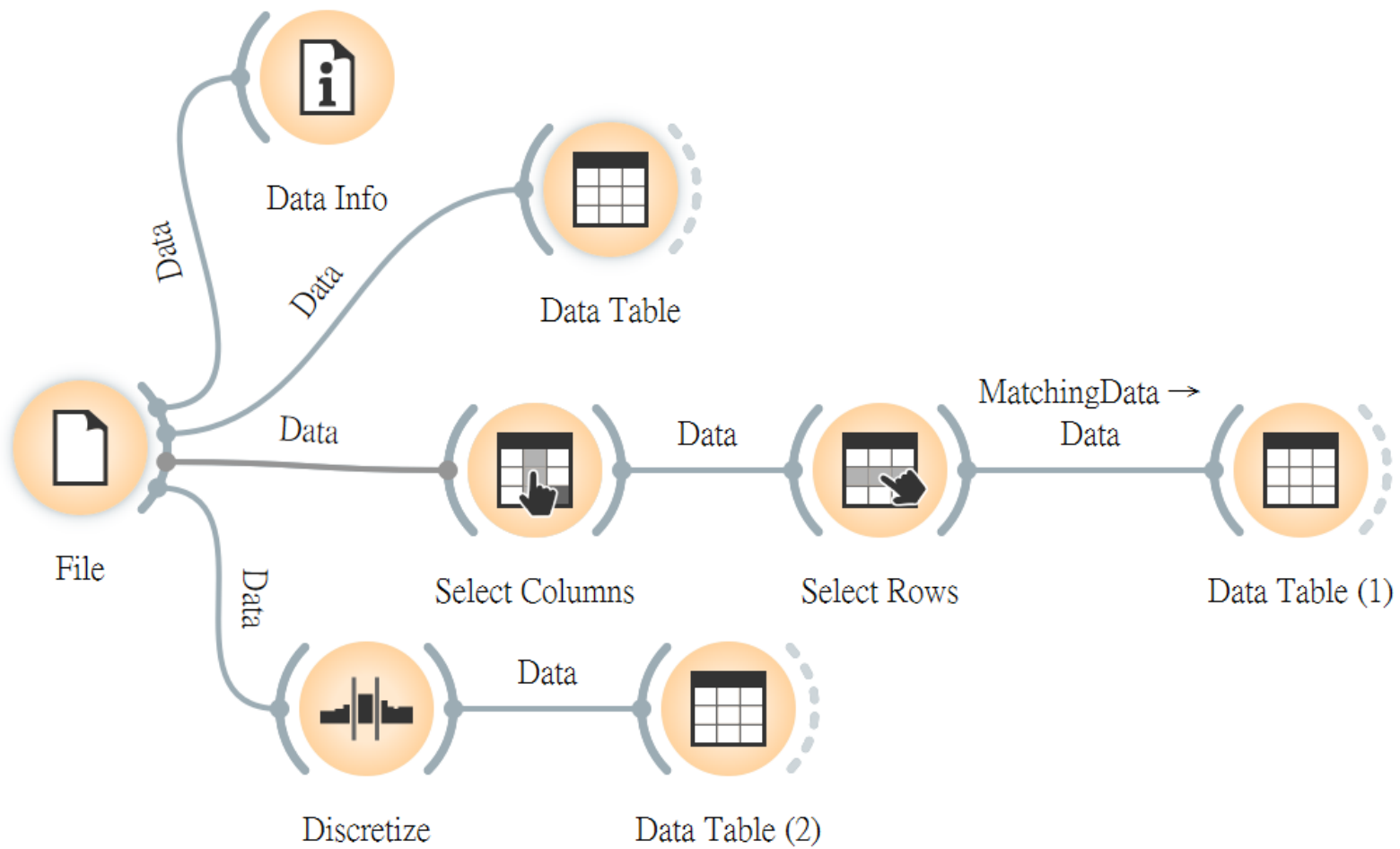
設定：Discretize元件



實際操作

- 依照下一頁建立資料流程圖
- File元件導入Score.csv
- 操作設定Discretize元件
 - 設定不同分組方式，並觀察其結果差異。
 - 設定不同組數，並觀察其結果差異。
 - 搭配Select Columns、Select Rows元件操作，並觀察其結果差異。

建立資料流程圖





資料表操作：

自訂函數來操作行與列



學習目標

- 使用 Feature Constructor 元件

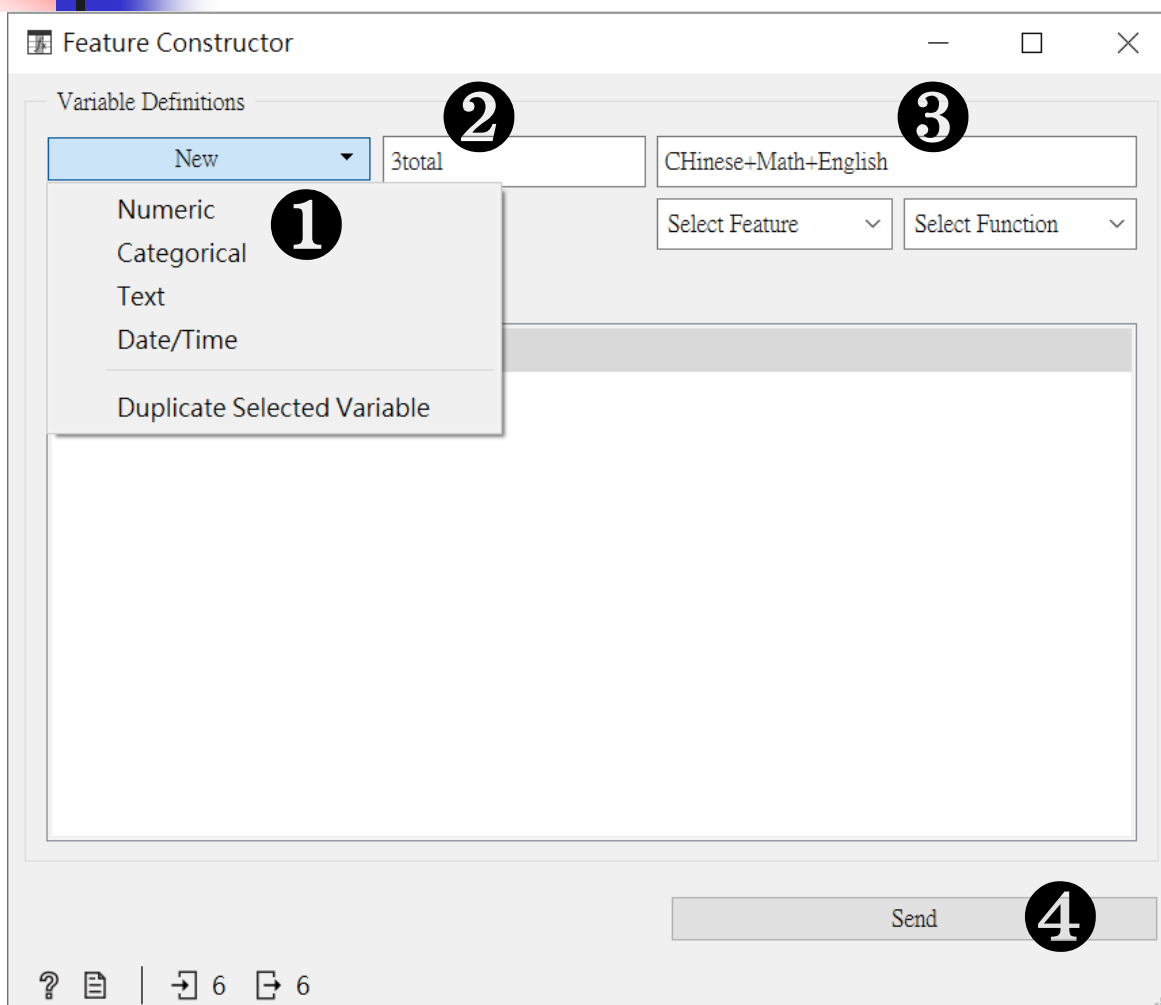


Feature Constructor 元件說明

Feature Constructor 是類似Excel公式的建構器

- ✓ 以”行”為單位，無法以單一儲存格操作
- ✓ 可以依照給定的欄位與公式建立新的資料值
- ✓ 內建函數：
 - 三角函數
 - 指數、對數 (c、log)
 - 特殊函數 (ex:取正負號的copysign、取最小整數值的ceil)
- ✓ 可以針對欄位進行四則運算
- ✓ 函數輸入格即是Python的單行程式碼輸入視窗，因此可以套用Python的語法 ex: sum(Chinese,Math,English)
或 使用 Column A + Column B + Column C

Feature Constructor 元件操作



① 選擇函數定義類別

② 輸入函式名稱

③ 輸入或選入函式內容

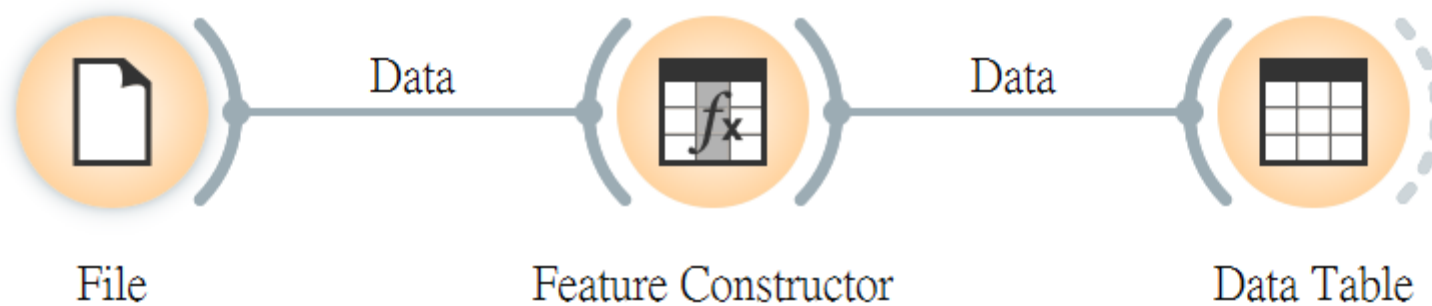
④ "Send" 傳送完成的所有設定內容



實際操作

- 依照下一頁建立資料流程圖
- File元件導入Score.csv
- 操作設定 Feature Constructor 元件
 - 建立新函式，使Data Table顯示增加一欄加總國語、數學、英語三科成績。
 - 以Python語法建立新函式，使Data Table顯示增加一欄加總五科成績，且數學、自然各加權50%。

建立資料流程圖





資料清理： 空資料與離群值



學習目標

- 使用空資料清除元件 Impute
- 瞭解離群值偵測元件 Outlier



Impute 元件說明

- Impute 元件用途在於清除空資料 (空值)
- 提供五種清除空資料方法：
 - 將平均值(Average)或眾數(Most frequent)填入空值
 - 平均值：當該行為連續型資料時
 - 眾數：當該行資料為離散型資料時
 - 給固定值(As a distinct values)
 - 使用 Simple Tree 根據其他列的資料建立模型，從既有的列數值中選擇最接近值來填入 (Model-based imputer)
 - 亂數 (Random value)
 - 移除該列 Remove instances with unknow values)
- 亂數 (Random value)

Impute 元件說明

Impute

五種移除空資料方法

Default Method

☒ Don't impute

☐ Average/Most frequent

☐ As a distinct value

☐ Model-based imputer (simple tree)

☐ Random values

☐ Remove instances with unknown values

Individual Attribute Settings

N Chinese

N Math

N English

N Science

N Society

欄位可個別
設定不同清
除空值方法

☒ Default (above)

☐ Don't impute

☐ Average/Most frequent

☐ As a distinct value

☐ Model-based imputer (simple tree)

☐ Random values

☐ Remove instances with unknown values

☐ Value

0.000

Restore All to Default

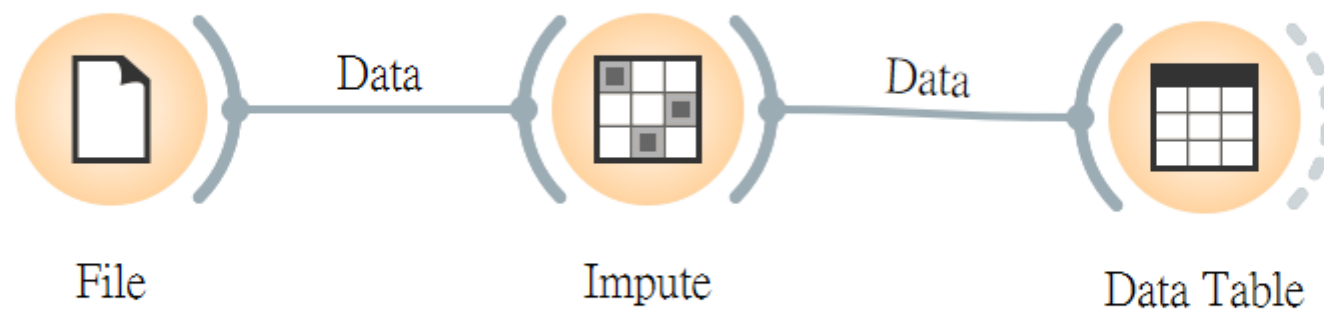
☒ Apply Automatically



實際操作

1. 依照下一頁建立資料流程圖
2. File元件導入Score.csv
3. 使用 Imputer 元件清除空資料
 - 開啟 Imputer 元件
 - 選擇使用五種方法清除空值
 - 使用Data Table觀察結果及五種結果的差異

建立資料流程圖





Outlier 元件說明

- Orange 3 中離群值採用分離器區分“離群”及“不離群”
- 離群值採用兩種機器學習算法：
 - 一階支持向量機(One class SVM with non-linear kernel)
 - 協方差估計 (Covariance estimator)



Thank you!!
