


图 52

## 3.4.2 预处理

### 3.4.2.1 缺失值处理

图标: 

**描述:** 缺失值处理是数据预处理的一部分, 由于采集的数据存在一些属性值的缺省, 如果不做处理, 将直接影响后续算法的挖掘效果, 严重时甚至得到错误的结果。处理方法有删除缺失值、中位数插补、众数插补、均值插补、线性插值、多项式插值。

#### 字段属性

**特征列:** 必选。选用中位数插值法、众数插值法、均值插值法时, 请选择数值型数据, 如果勾选了非数值类型数据, 则会自动过滤, 下个组件可能无法获取所有列。勾选的列将传入下一个组件。

#### 参数设置

**处理方式:** 选择对该缺失数据列的处理方式, 可以选择删除缺失值、中位数插补、众数插补、均值插补、线性插值、多项式插值。

#### 输出

**表结果:** 缺失值处理结果。

**报告:** 无。

#### 示例

下面对某数据进行缺失值处理。

- 勾选需要进行缺失值处理的数据, 将会在勾选的数据内查找缺失值, 并进行相应的处理。如图 53、**错误!未找到引用源。** 所示。
- 选择处理方式为【**删除法**】, 如图 54 所示。
- 运行该组件后可通过查看数据查看结果。

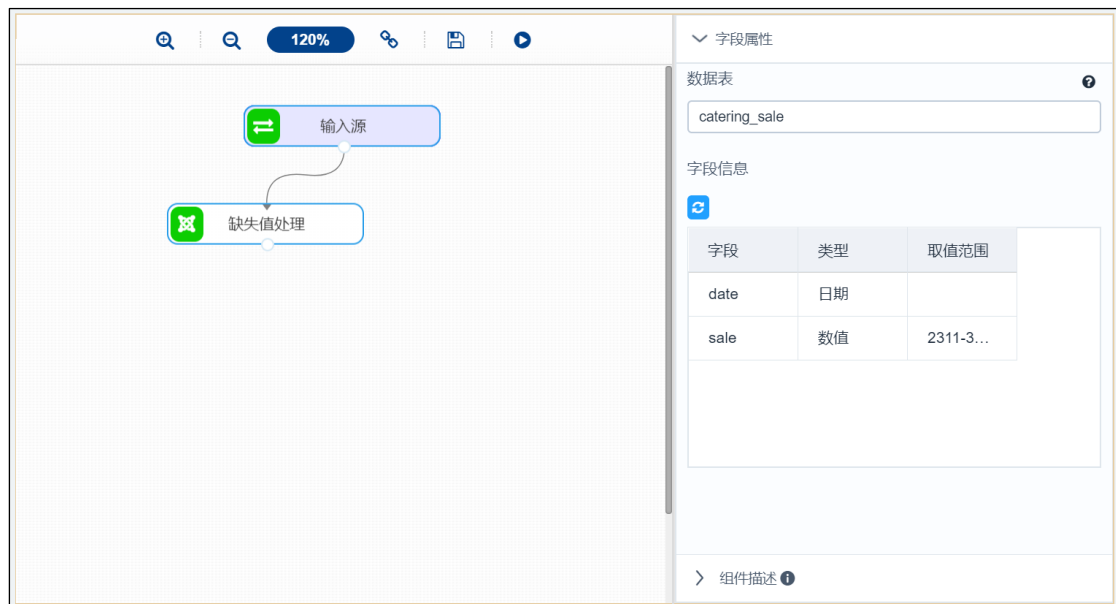


图 53

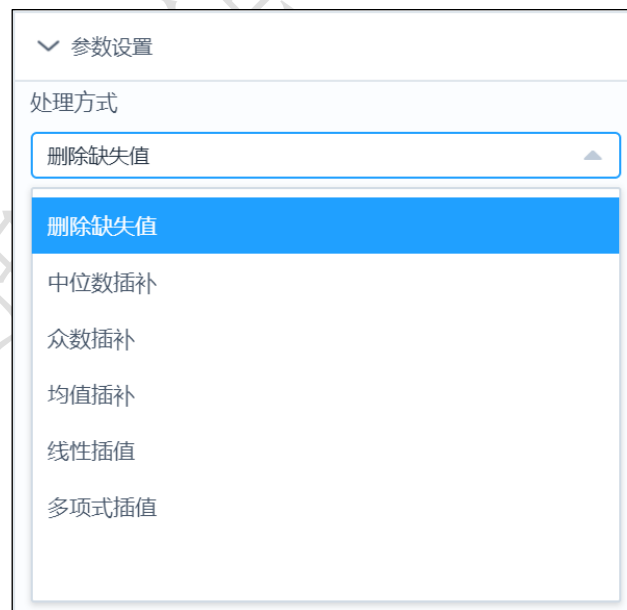


图 54

### 3.4.2.2 记录选择

图标:  记录选择

描述: 记录选择是对数据表的行进行筛选, 只留下满足条件的数据行。

字段属性

特征列：选择需要进行记录选择的列，勾选的列将传入下一个组件。如图 55 所示。



图 55

### 参数设置

参数设置包括：过滤器的增加和删除、刷新列、运算符、过滤列、过滤条件、过滤值，如图 56 所示。

**添加 (+) 和删除**：通过点击添加按钮添加一个列的过滤设置，通过删除图标减少一个列的过滤设置

**刷新列**：想要获取数据，则必须事先通过点击刷新按钮

**运算符**：提供各条件之间 and 和 or 的选择，

**过滤列**：获取上级操作单元节点的列信息，供用户选择(单选)

**过滤条件**：目前操作符支持 “=”，“!=”，“>”，“<”，“>=” 和 “<=”

**过滤值**：条件值选择，当过滤值为字符类型时，需添加双引号



图 56

### 输出

表结果：记录选择结果。

报告：无。

### 示例

下面对某数据进行记录选择，数据一共包括四个字段：id、r、f、m。选择满足  $r > 27$  的所有数据。

- 勾选需要进行记录选择的特征列，如图 57 所示。
- 依次点击【加号】及【刷新】按钮，选择字段、运算符、值。如图 58 所示。
- 运行该组件，对组件右击，选择查看数据，结果如图 59 所示。

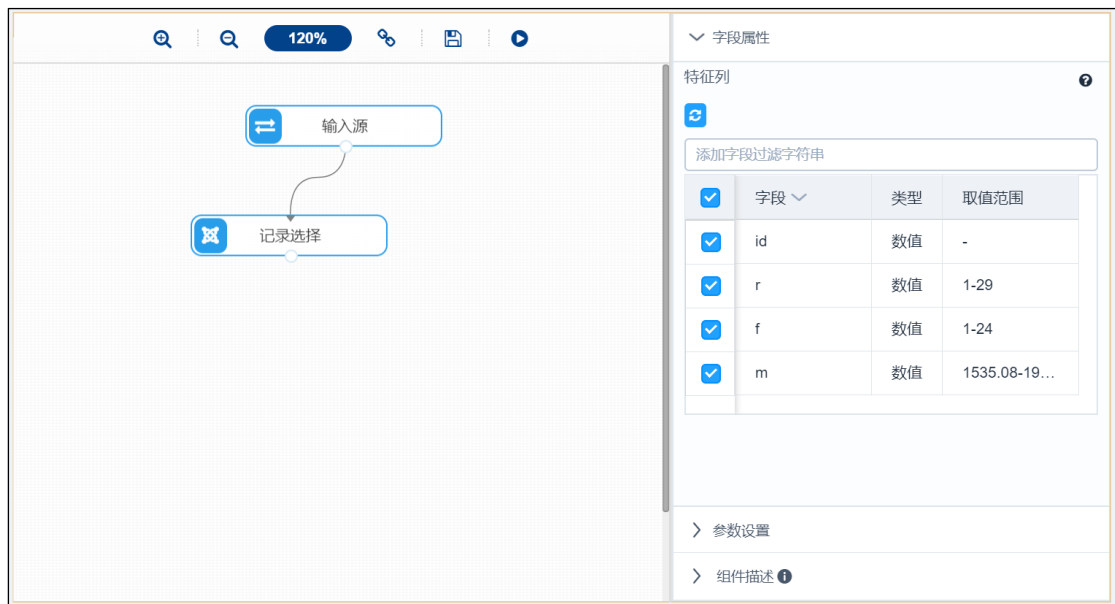


图 57

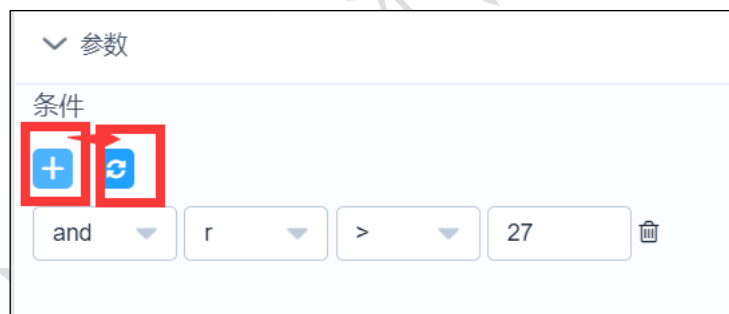


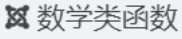
图 58

预览数据 (仅显示前100条)

id	f	m	r
14	16	1957.44	30
17	2	1016.34	93
26	21	1628.68	30
30	7	5318.81	60
52	8	1865.99	30
53	8	1791.44	28
58	4	2920.81	66
77	11	1461.63	78
81	2	227.14	28

图 59

### 3.4.2.3 数学类函数

图标: 

**描述:** 数学类函数是对勾选的某列运用数学运算函数。

#### 字段属性

**特征列:** 必须包含待运算的列, 勾选的列将传入下一个组件。

#### 参数设置

**待运算的列:** 请选择数值型数据。

**运算函数:** 包括向上取整、绝对值、向下取整、平方根、返回整数。

#### 输出

**表结果:** 运算结果。

**报告:** 无。

#### 示例

下面对某数据的一列进行向上取整, 原数据如图 60 所示。



sepal_length	sepal_width	petal_length	petal_width	species
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa

图 60

- 在特征列中勾选需要传入下个组件的数据, 必须包含待运算的列。如图 61 所示。



图 61

- 打开参数设置选项卡，在【运算函数】下拉框中选择【向上取整】，如图 62 所示。

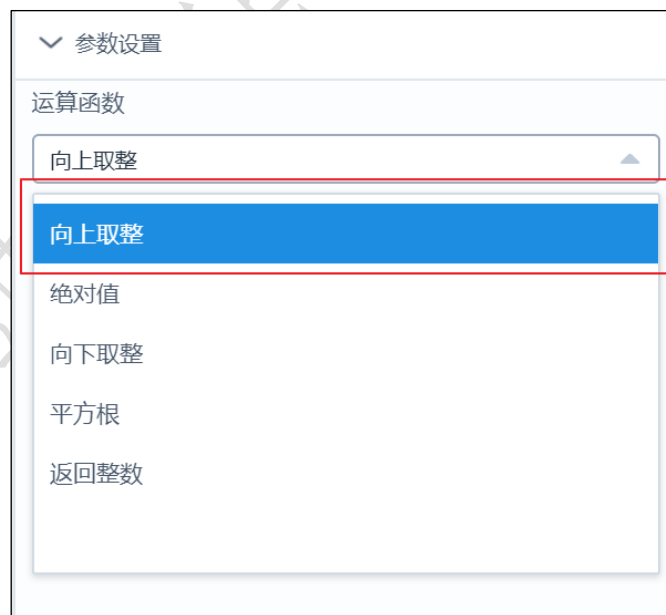


图 62

- 运行成功后,点击查看数据,结果如图 63 所示.

预览数据	
sepal_width	petal_length
4	1.4
3	1.4
4	1.3
4	1.5
4	1.4
4	1.7
4	1.4
4	1.5

共 150 条 25 条/页 < 1 2 3 4 5 6 > 前往 1 页

图 63

#### 3.4.2.4 表合并

图标: 

**描述:** 表合并是指两张表通过行或列合并成一张表, 不需要关键字段。

**字段属性**

**左表特征列:** 勾选左表需要合并的列, 需要注意的是左表特征列与右表特征列列名不能重合。

**右表特征列:** 勾选右表需要合并的列, 需要注意的是左表特征列与右表特征列列名不能重合。如图 64 所示。





图 64

### 参数设置

合并方式：必选。可选择行合并，列合并。需要注意的是，选择列合并时，两个表的行数需要一样；选择行合并时，两个表的列数需要一样，如图 65 所示。

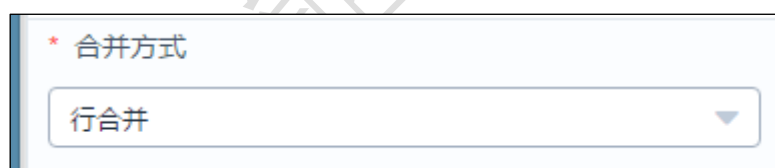


图 65

### 输出

表结果：合并结果。

报告：无。

### 示例

下面按列将两个表合并为一个表。

- 勾选需要进行合并的字段，左表选择 f, m 两列，右表选择 r 列，如图 66 所示。
- 合并方式选择列合并，如图 67 所示。
- 运行该组件，右击选择查看数据，结果如图 68 所示。

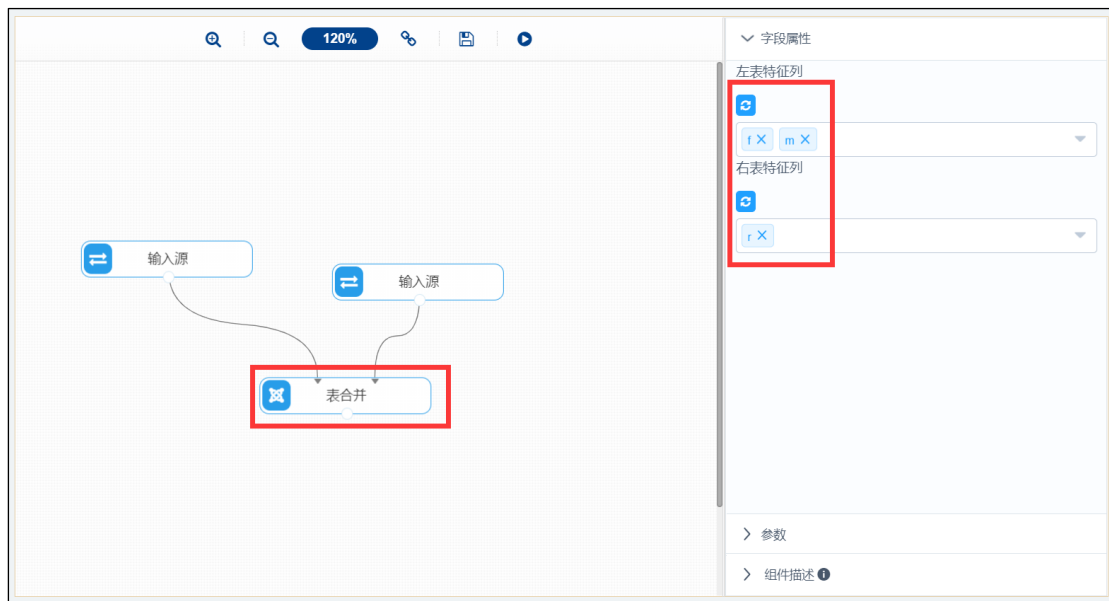


图 66

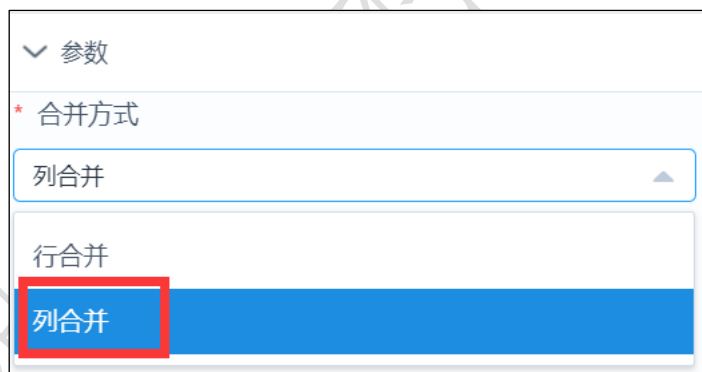


图 67

f	m	r
6	232.61	27
5	1507.11	3
16	817.62	4
11	232.81	3
7	1913.05	14
6	220.07	19
2	615.83	5

图 68

### 3.4.2.5 表连接

图标: 

**描述:** 表连接是指两张表通过某列进行关联, 合成一张表。

#### 字段属性

**左表:** 必选。选择左表需要关联的列, 必须包含左表连接关键字列。

**右表:** 必选。选择右表需要关联的列, 必须包含右表连接关键字列, 如图 69 所示。



字段属性

\* 左表特征列 ?

id X m X

\* 右表特征列 ?

id X r X

左表连接关键字 ?

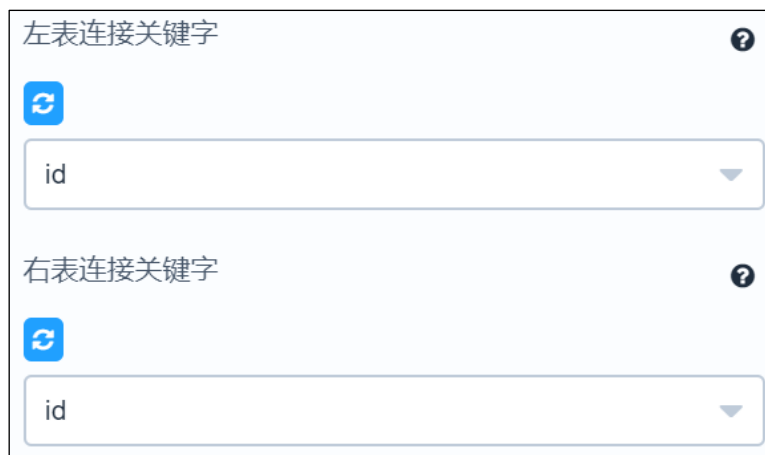
图 69

#### 参数设置

**左表连接关键字:** 必选。选择和右表关联的列。

**右表连接关键字:** 必选。选择和左表关联的列。

**连接方式:** 必选。支持左外连接, 内连接, 右外连接, 全外连接。其中左外连接返回左表中所有的记录以及右表中连接字段相等的记录; 右外连接返回右表中所有的记录以及左表中连接字段相等的记录; 内连接返回两个表中连接字段相等的记录; 全外连接返回两个表中的记录, 如图 70、图 71 所示。



左表连接关键字

刷新

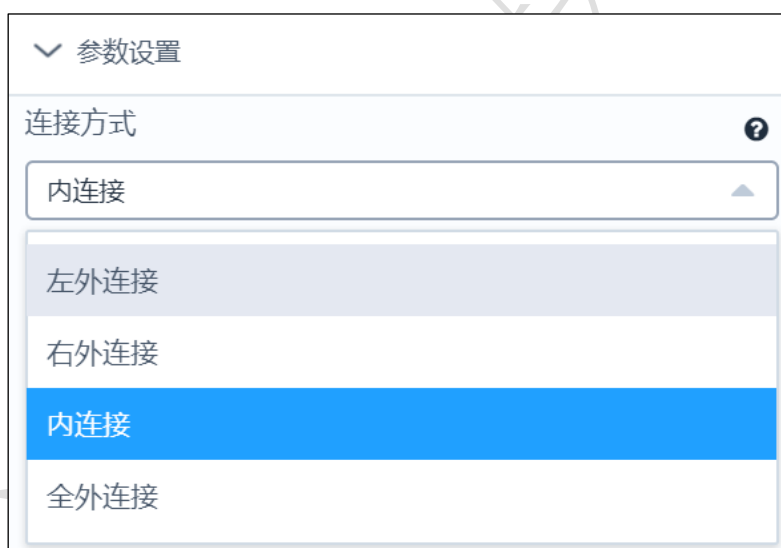
id

右表连接关键字

刷新

id

图 70



参数设置

连接方式

内连接

左外连接

右外连接

内连接

全外连接

图 71

### 输出

表结果：表连接结果。

报告：无。

### 示例

下面对两个数据表进行表连接。

- 勾选需要进行连接的字段，左表及右表的特征列必须包含连接需要用到的连接关键字。该数据关键字为 id，左边选择 id，m 两列，右表选择 id，f 两列。如图 72 所示。
- 勾选左表的关键列为 id，右表的关键列为 id，使用全外连接。如图 73 所示。

- 运行该组件，右击选择查看数据，结果如图 74 所示。



图 72



图 73

id	m	f
1	232.61	6
2	1507.11	5
3	817.62	16
4	232.81	11
5	1913.05	7
6	220.07	6
7	615.83	2

图 74

### 3.4.2.6 平稳性检验

图标:

平稳性检验

**描述:** 平稳性检验是为了确定序列是否存在确定趋势，否则将会产生“伪回归”问题。伪回归是说，有时数据的高度相关仅仅是因为二者同时随时间有向上或向下的变动趋势，并没有真正联系。这样数据中的趋势项，季节项等无法消除，从而在残差分析中无法准确进行分析。

**字段属性**

**时序列:** 必选。选择想要进行检验的数据列，请选择数值型数据，如果该列数据含有缺失值，则会自动删除，如图 75 所示。



图 75

**参数设置**

无

**输出**

表结果：无。

报告：Test statistic、p-value、Number of lags used、Number of observations used for the ADF regression and calculation of the critical values、Critical values for the test statistic at the 5 %、Critical values for the test statistic at the 1 %、Critical values for the test statistic at the 10 %、自相关图。

**示例**

下面对某列数据进行平稳性检验。

- 选择时序列，数据必须为数值型。如图 76 所示。
- 运行该组件，对组件右击，选择查看报告，结果如图 77 所示。

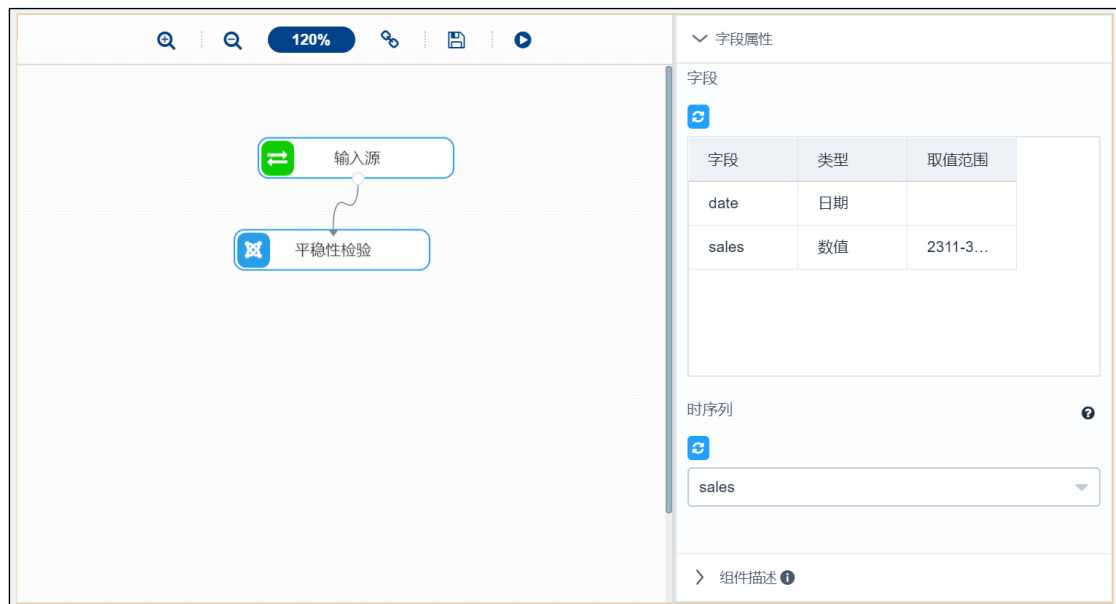


图 76



图 77

### 3.4.2.7 纯随机性检验

图标:  随机性检验

**描述:** 纯随机性检验又称为白噪声检验,是专门用来检验序列是否为纯随机序列的一种方法。纯随机序列的序列值之间没有任何相关关系,也就是没有什么统计规律可言,各项之间也就没有任何关联,这样的序列没有挖掘的意义。

字段属性



字段属性包括：字段信息、待检验序列，如图 78 所示。

待检验序列：必选。选择想要进行检验的列，请数值型数据。

The screenshot shows a configuration window titled '字段属性' (Field Properties). It contains two main sections:

- 字段 (Fields):** A table with columns '字段' (Field), '类型' (Type), and '取值范围' (Value Range).

字段	类型	取值范围
date	日期	
sales	数值	2311-3...
- 待检验序列 (Sequence to be tested):** A dropdown menu with a refresh icon and a help icon. The selected value is 'sales'.

图 78

### 输出

表结果：对应各滞后阶数的 P 值。

报告：无。

### 示例

下面对某列数据进行纯随机性检验。

- 选择待检验序列，数据必须为数值型。如图 79 所示。
- 运行成功后，选择查看数据，如图 80 所示。

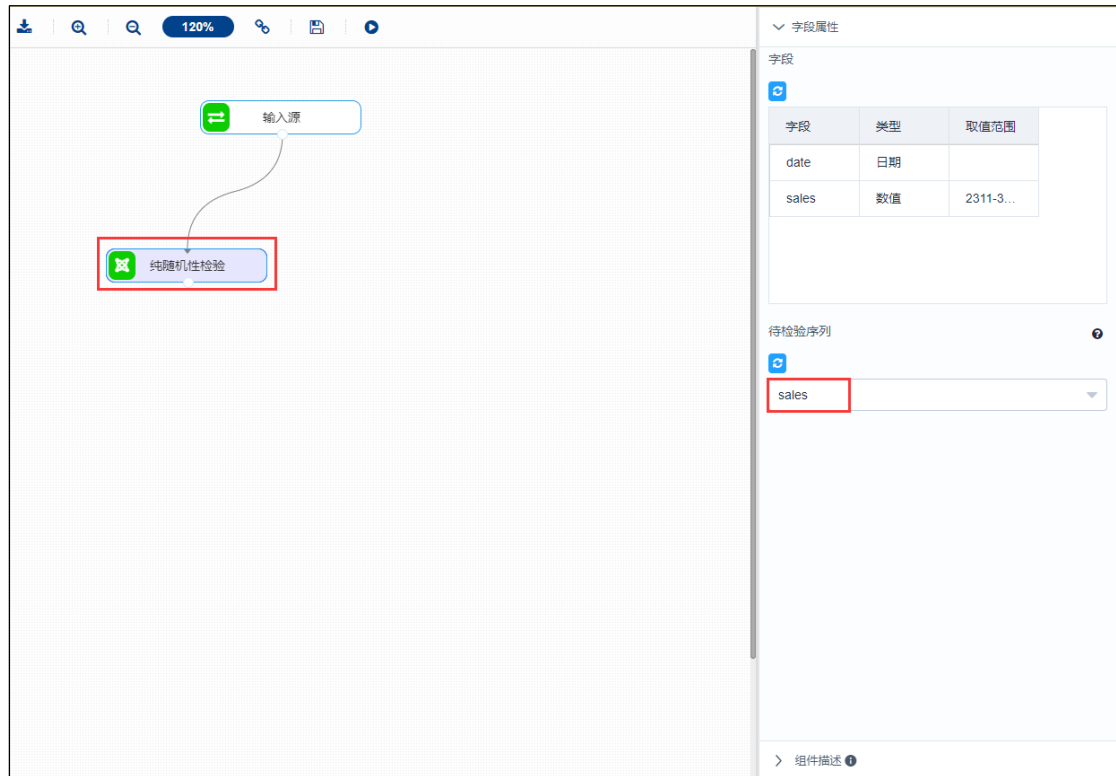


图 79

预览数据	
lags	pvalue
1	0.05664850864089642
2	0.08171417014241165
3	0.11638329840952293
4	0.2064018913330824
5	0.3073699633365428
6	0.3672622868506844
7	0.4476144801344415
8	0.27237256179528074

共 40 条    25 条/页    < 1 2 >    前往 1 页

图 80

### 3.4.2.8 记录去重

图标: 

**描述:** 记录去重是去除数据表中的重复的行数据，只保留其中一行数据。

#### 字段属性

**特征列:** 必选。选择需要进行去重的列，勾选的字段可传入下个组件，如图 81 所示。



图 81

#### 参数设置

无

#### 输出

表结果: 去重后的数据。

报告: 无。

#### 示例

- 勾选需要去重的数据，勾选的列将传入下个组件。如图 82 所示。
- 运行成功后，可右键查看数据，结果如图 83 所示。

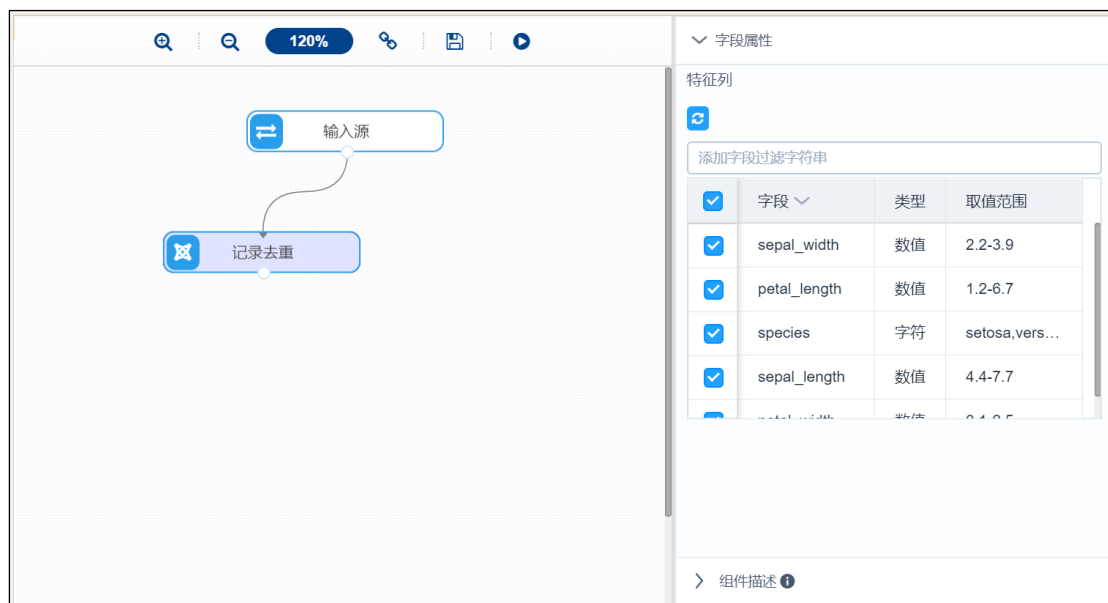


图 82

The screenshot shows a '预览数据' (Preview Data) window. It displays a table with the following data:

sepal_width	petal_length	species	sepal_length	petal_width
3.5	1.4	setosa	5.1	0.2
3	1.4	setosa	4.9	0.2
3.2	1.3	setosa	4.7	0.2
3.1	1.5	setosa	4.6	0.2
3.6	1.4	setosa	5	0.2
3.9	1.7	setosa	5.4	0.4
3.4	1.4	setosa	4.6	0.3
3.4	1.5	setosa	5	0.2

At the bottom of the window, it shows '共 149 条' (Total 149 rows), '25 条/页' (25 rows per page), and a pagination control with page numbers 1 through 6, and a '前往' (Go to) button.

图 83

### 3.4.2.9 数据离散化

图标:  数据离散化

**描述:** 某些模型算法，特别是某些分类算法如 ID3 决策树算法和 Apriori 算法等，要求数据是离散的，此时就需要将连续型特征（数值型）变换成离散型特征（类别型），即连续特征离散化。常用的离散化方法主要有三种：等宽法，等频法和通过聚类分析离散化（一维）。

**字段属性**

**待离散化数据:** 必选。请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，

下个组件可能无法获取所有列。勾选多列时，自动对每一列数据进行离散化如图 84 所示。

字段	类型	取值范围
a	数值	-
b	数值	-

\* 待离散化数据

a × b ×

图 84

### 参数设置

离散化方式：选取要使用的离散方式，支持等宽、等频、聚类离散化，默认等宽。

离散个数：离散的个数，默认 2，如图 85 所示。

离散化方式

等宽

离散个数

3

图 85

### 输出

表结果：对勾选的每一列进行离散化后的结果。

报告：无。

### 示例

下面对数据进行离散化。原数据如图 86 所示。

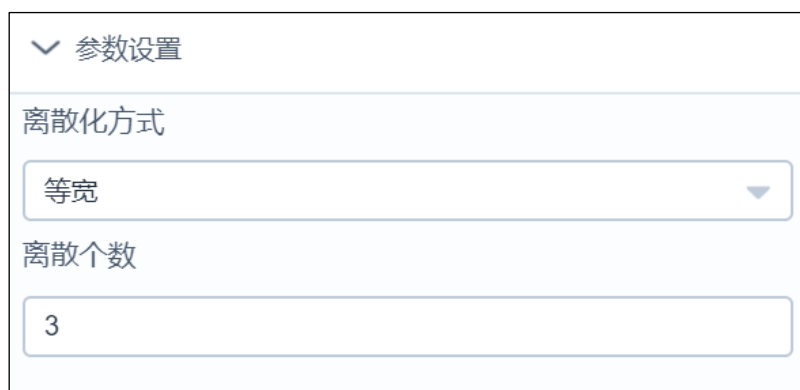
预览数据	
a	b
1	2
2	3
3	4
4	5
5	6

图 86

- 勾选需要进行离散化的数据，如图 87 所示。
- 选择离散化方式为等宽，离散个数为 2。如图 88 所示。
- 运行该组件，右击选择查看数据，如图 89 所示。



图 87



参数设置

离散化方式

等宽

离散个数

3

图 88



预览数据	
a	b
(0.995, 2.333]	(1.995, 3.333]
(0.995, 2.333]	(1.995, 3.333]
(2.333, 3.667]	(3.333, 4.667]
(3.667, 5.0]	(4.667, 6.0]
(3.667, 5.0]	(4.667, 6.0]

图 89

### 3.4.2.10 排序

图标: 

**描述:** 根据某一列的顺序将所有数据重新排序。

#### 字段属性

**特征列:** 勾选的列必须包含关键字段。勾选的列将传入下一个组件。如图 90 所示。

**关键字:** 必选，由该列值的顺序将待排序的数据重新排序。如图 91 所示。



<input checked="" type="checkbox"/>	字段	类型	取值范围
<input checked="" type="checkbox"/>	a	数值	-
<input checked="" type="checkbox"/>	b	数值	-

图 90

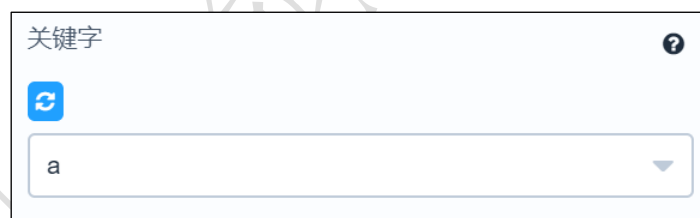


图 91

### 参数设置

排序方式：将数据按某列的顺序将所有数据按照升序或降序重新排序，如图 92 所示。



图 92

### 输出

表结果：按照某列排序后的数据。



报告：无。

## 示例

下面将按列 m 的值对数据进行升序排序。

- 勾选需要进行排序的特征列，选择关键列 m，特征列必须包含关键列。如图 93 所示。
- 选择升序。如图 94 所示。
- 运行该组件，结果如图 95 所示。



图 93

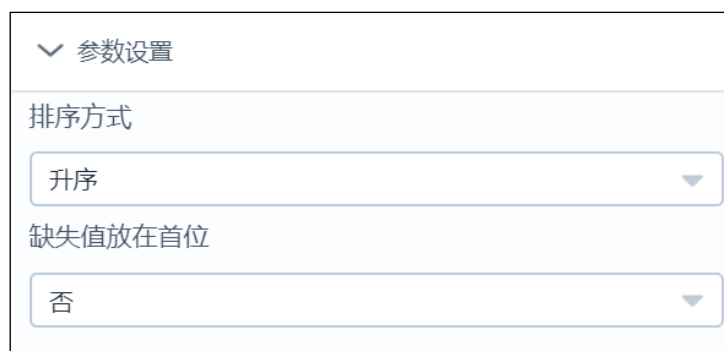


图 94

预览数据			
id	m	r	f
848	23.24	4	1
773	28.43	20	1
915	33.58	17	1
445	45.23	7	1
813	49.61	16	1
776	49.7	4	1
294	50.14	3	2
129	53	28	1

共 940 条 25 条/页 < 1 2 3 4 5 6 ... 38 > 前往 1 页

图 95

### 3.4.2.11 数据拆分

图标: 

**描述:** 数据拆分对全量数据进行简单随机抽样, 将数据拆分为训练数据和测试数据。

**字段属性**

**特征列:** 必选。选择进行拆分的数据, 勾选的列将传入下个组件, 如图 96 所示。

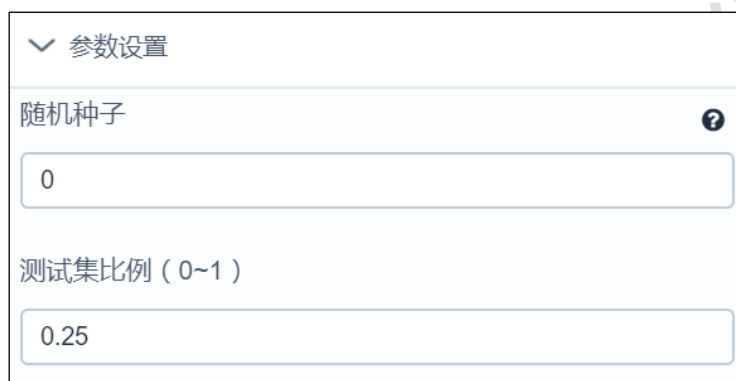
特征列			
<input checked="" type="checkbox"/>	字段	类型	取值范围
<input checked="" type="checkbox"/>	sepal_width	数值	2.2-3.9
<input checked="" type="checkbox"/>	petal_length	数值	1.2-6.7
<input checked="" type="checkbox"/>	species	字符	setosa,vers...
<input checked="" type="checkbox"/>	sepal_length	数值	4.4-7.7
<input checked="" type="checkbox"/>	petal_width	数值	0.4-0.5

图 96

## 参数设置

随机种子：可以理解为一个序号，这个序号交给一个数列管理器，通过这个序号，从管理器中取出一个数列，这个数列就是通过那个序号得到的随机数。

训练集比例：设置训练数据的比例，范围在 0-1 之间，默认 0.25，如图 97 所示。



参数设置

随机种子 ?

0

测试集比例 (0~1)

0.25

图 97

## 输出

表结果：训练集与测试集。

报告：无。

## 示例

下面将某数据拆分为训练集、测试集。

- 勾选需要进行数据拆分的数据，如图 98 所示。
- 保留默认的随机种子，设置测试集比例为 0.25，如图 99 所示。
- 运行该组件，选择查看数据，查看对应的数据集，如图 100 所示。



图 98

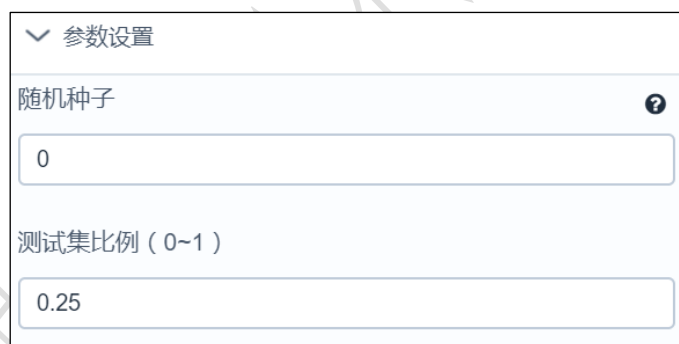


图 99



图 100

### 3.4.2.12 频数统计

图标: 

描述: 频数统计对某种特征的数(标志值)出现的次数进行统计。

字段属性

待统计列: 需要计算频数的一列数据。



配置界面显示“字段属性”下的“待统计列”选择框，当前选中了“species”。

参数设置

无

输出

表结果: 频数表。

报告: 无。

示例

对某列数据进行频数统计，数据如图 101 所示。

预览数据				
sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa

共 150 条 25 条/页 < 1 2 3 4 5 6 > 前往 1 页

图 101

- 勾选需要进行频数统计的数据，如图 102 所示。

- 运行该节点，右击选择查看数据，如图 103 所示。

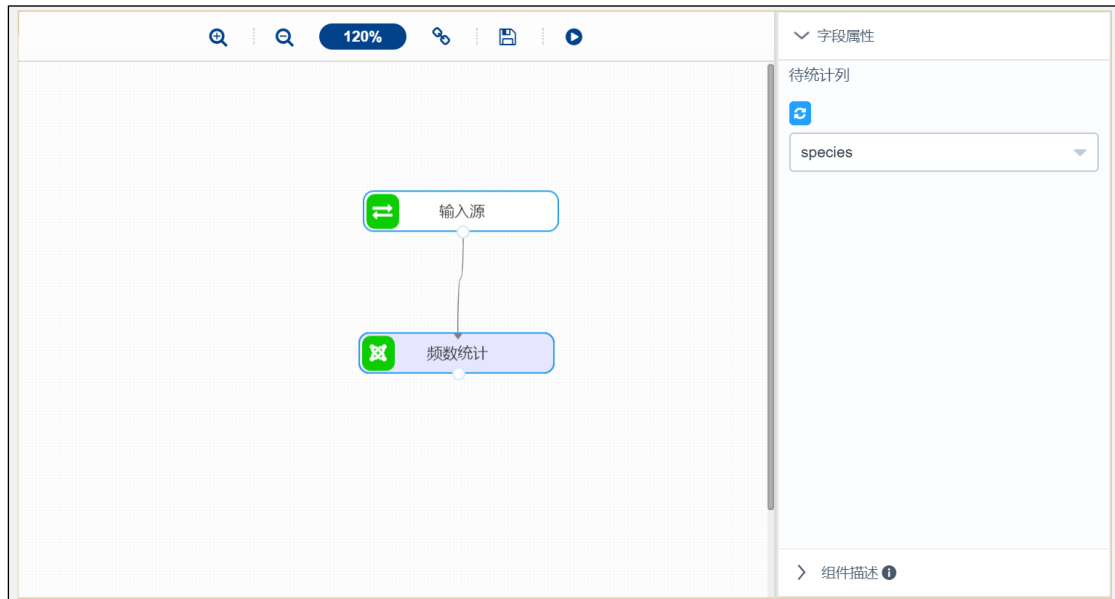


图 102

预览数据	
species	count
setosa	50
versicolor	50
virginica	50

图 103

### 3.4.2.13 新增序列

图标: 

**描述:** 新增序列是指在原有数据基础上，新增一列自增序列，作为标识或其他特定作用。

**字段属性**

**特征列:** 增加的序列会在勾选的字段基础上添加。如图 104 所示。



图 104

### 参数设置

新增序列列名：定义新增序列的列名。默认 new，如图 105 所示。



图 105

### 输出

表结果：新增序列后的表。

报告：无。

### 示例

下面对某数据新增一列自增序列，原表共有四个字段：id, f, m, r。需要新增一列列名为 new 的自增序列。

- 勾选原表字段，新增序列需要在勾选的数据基础上新增。如图 106 所示。
- 保留定义的新增序列列名为 new。如图 107 所示。
- 结果如图 108 所示。

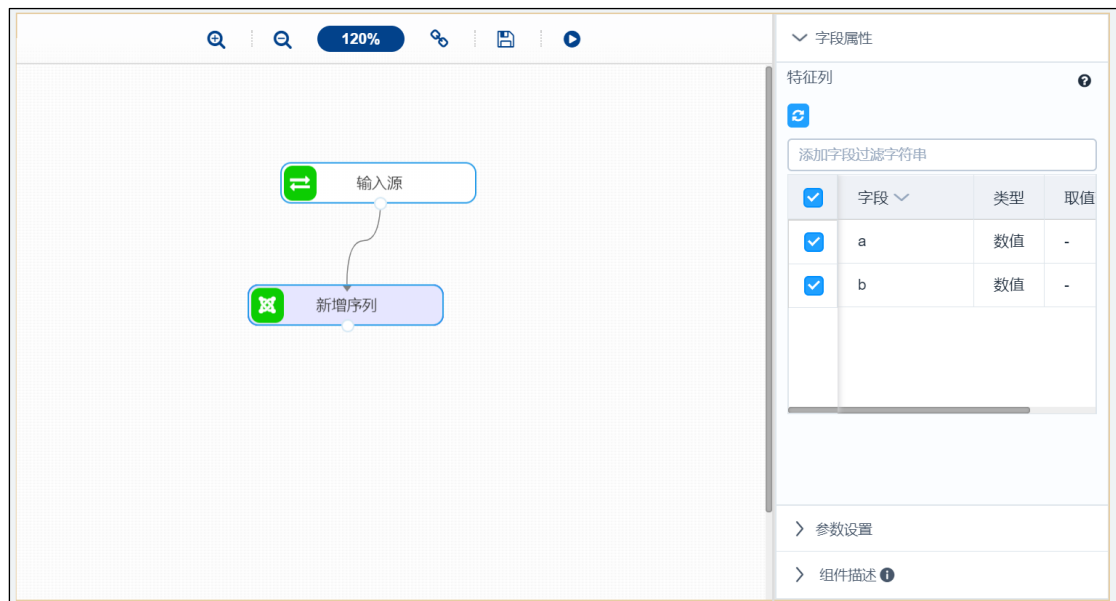


图 106



图 107

预览数据

new	a	b
1	1	2
2	2	3
3	3	4
4	4	5
5	5	6

图 108

### 3.4.2.14 K 步差分

图标: 



**描述：**K 步差分是指相距 k 期的两个序列值之间的减法运算。

### 字段属性

待差分序列：请选择数值型数据，如图 109 所示。

字段	类型	取值范围
sepal_width	数值	2.2-3.9
petal_length	数值	1.2-6.7
species	字符	setosa,...
sepal_length	数值	4.4-7.7

待差分序列

sepal\_width

图 109

### 参数设置

K 期：整数型，默认 1，如图 110 所示。

参数设置

K期

1 - +

图 110

### 输出

表结果：差分结果。

报告：无。

### 示例

下面对某数据进行 1 步差分。原数据如图 111 所示。

预览数据	
a	b
1	2
2	3
3	4
4	5
5	6

图 111

- 选择待差分序列，如图 112 所示。
- 设置差分步数 K，如图 113 所示。
- 运行成功后，选择查看数据，结果如图 114 所示。

The screenshot shows a workflow editor with two components: '输入源' (Input Source) and 'K步差分' (K-step Difference). The right-side panel displays the '字段属性' (Field Properties) section, which includes a table of field attributes and a dropdown menu for selecting the '待差分序列' (Sequence to be Differenced).

字段	类型	取值范围
a	数值	-
b	数值	-

待差分序列: b

参数设置

组件描述

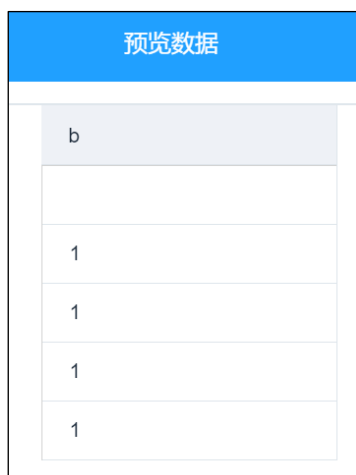
图 112

参数设置

K期

1 - +

图 113



b
1
1
1
1

图 114

### 3.4.2.15 分组聚合

图标:  分组聚合

**描述:** 分组聚合是指将数据按照某个键拆分为多组, 使用一个函数应用到各个分组上产生一个新值, 最后将执行的结果合并。

#### 字段属性

待分组与聚合的列: 当聚合方式为 `count` 外, 请选择数值型数据。需要注意的是勾选的列不包含键, 如图 115 所示。



图 115

键: 按照该列的值将数据分组, 如图 116 所示。

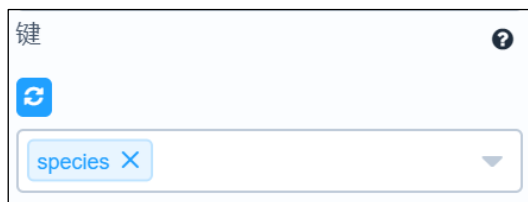


图 116

### 参数设置

聚合方式：应用到每个分组上的函数。包括 count、max:、mean、median、size、min、std:、sum。

### 输出

表结果：分组聚合结果。

报告：无。

### 示例

下面对某数据进行分组聚合,原数据如图 117 所示.

预览数据				
sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa

共 150 条 25 条/页 < 1 2 3 4 5 6 > 前往 1 页

图 117

- 勾选待分组与聚合列,将这些数据按照列 species 分组。如图 118 所示。
- 选择聚合方式为 max, 如图 119 所示。
- 运行成功, 选择查看数据, 如图 120 所示。

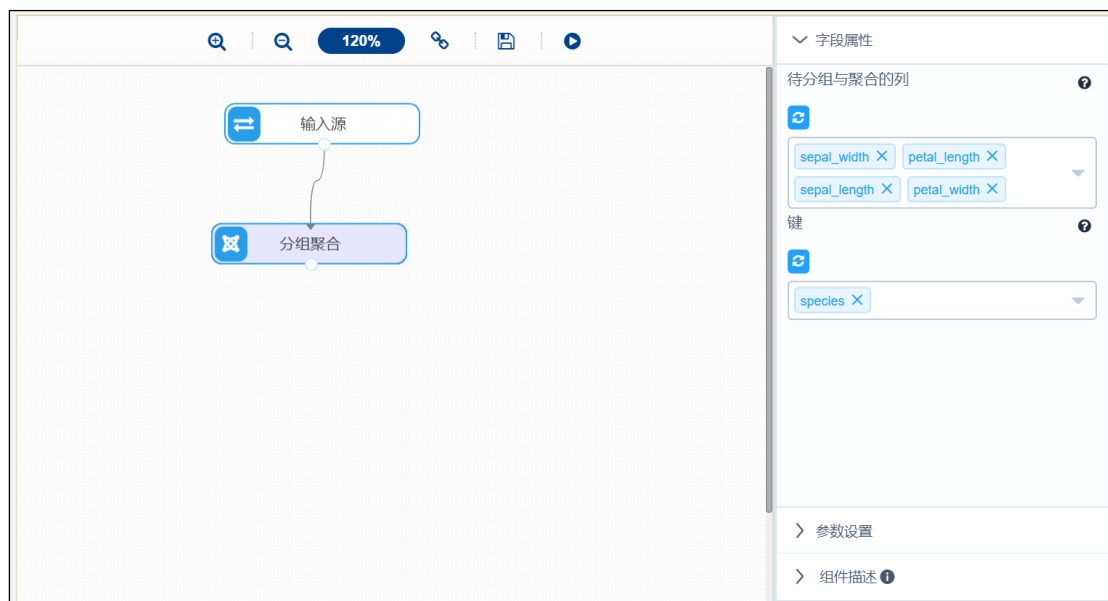


图 118

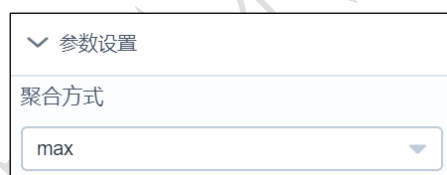


图 119

预览数据				
species	sepal_width	petal_length	sepal_length	petal_width
setosa	4.4	1.9	5.8	0.6
versicolor	3.4	5.1	7	1.8
virginica	3.8	6.9	7.9	2.5

图 120

### 3.4.2.16 数据标准化

图标:  数据标准化

**描述:** 数据标准化处理是将数据按比例缩放，使之落入一个小的特定区间。

#### 字段属性

**特征列:** 选择进行标准化的列，请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列，如图 121 所示。



图 121

### 参数设置

标准化方式：标准化方式包括极差标准化、零均值标准化和小数定标标准化，默认零均值标准化。

最小值：选择极差标准化时有效。

最大值：选择极差标准化时有效。如图 122 所示



图 122

### 输出

表结果：标准化结果。

报告：无。

### 示例

下面对某数据进行标准化处理。原数据如图 123 所示。

预览数据			
id	r	f	m
1	27	6	232.61
2	3	5	1507.11
3	4	16	817.62
4	3	11	232.81
5	14	7	1913.05
6	19	6	220.07
7	5	2	615.83
8	26	2	1059.66

共 940 条 25 条/页 < 1 2 3 4 5 6 ... 38 > 前往 1 页

图 123

- 勾选需要进行数据标准化的数据。如图 124 所示。
- 选择标准化方式为零均值标准化，如图 125 所示。
- 结果如图 126 所示。

The screenshot shows a data processing workflow in a software interface. The main workspace contains two components: '输入源' (Input Source) and '数据标准化' (Data Standardization), connected by a flow line. The right-hand panel, titled '字段属性' (Field Properties), displays a table of selected features for standardization.

特征列	字段	类型	取值
<input type="checkbox"/>	id	数值	-
<input checked="" type="checkbox"/>	m	数值	1535
<input checked="" type="checkbox"/>	r	数值	1-29
<input checked="" type="checkbox"/>	f	数值	1-24

Below the table, there are sections for '参数设置' (Parameter Settings) and '组件描述' (Component Description).

图 124

∨ 参数设置

标准化方式

零均值标准化

最小值 ?

最大值 ?

图 125

预览数据		
m	r	f
-1.159328122362407	0.7645931616651589	-0.49384157634634734
0.62285862671049	-1.025302213157502	-0.6304144159360991
-0.34128412858042545	-0.9507232392065578	0.8718868195511712
-1.1590484539827959	-1.025302213157502	0.18902262160241196
1.19050153680751	-0.20493349969711572	-0.3572687367565955
-1.1768633297640345	0.16796137005760528	-0.49384157634634734
-0.6234555401892288	-0.8761442652556135	-1.0401329347053547
-0.00282945557485135	0.6900141877142147	-1.0401329347053547

共 940 条 25 条/页 < 1 2 3 4 5 6 ... 38 > 前往  页

图 126

### 3.4.2.17 衍生变量

图标:  衍生变量

**描述:** 衍生变量是指将一系列或多列通过基本运算生成新列。

**字段属性**

**特征列:** 必选。选择进行衍生变量的列。请选择数值型数据,增加的序列会在勾选的字段



基础上添加，如图 127 所示。



图 127

### 参数设置

**变量名：**新增列的列名，输入要求：;1.英文开头;2.小写英文、数字、下划线;3.长度 1-10，默认名称是 new。

**表达式：**必填。目前只支持四则运算符：+，-，\*，/。如图 128 所示。

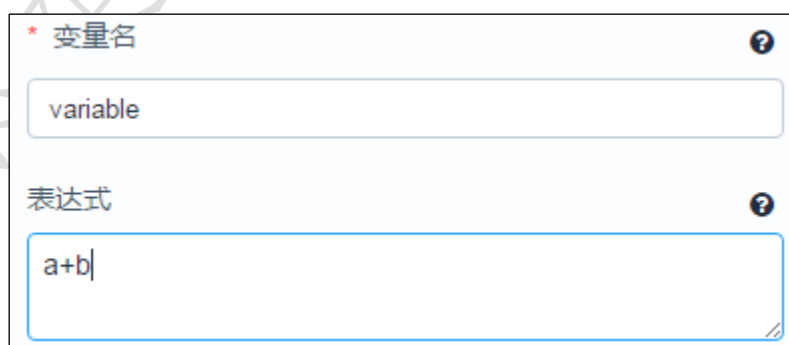


图 128

### 输出

**表结果：**勾选的列与衍生出来的列构成的表。

**报告：**无。

### 示例

下面对某数据使用衍生变量,将源数据 `sepal_width`,`petal_length` 两字段相加(`sepal_width + petal_length`) 构成新列 `new`。

- 勾选需要进行衍生变量的列。如图 129 所示。
- 定义新增列的列名为 `new`, 表达式为"`sepal_width + petal_length`", 如图 130 所示。
- 运行该节点, 右击选择查看数据, 如图 131 所示。

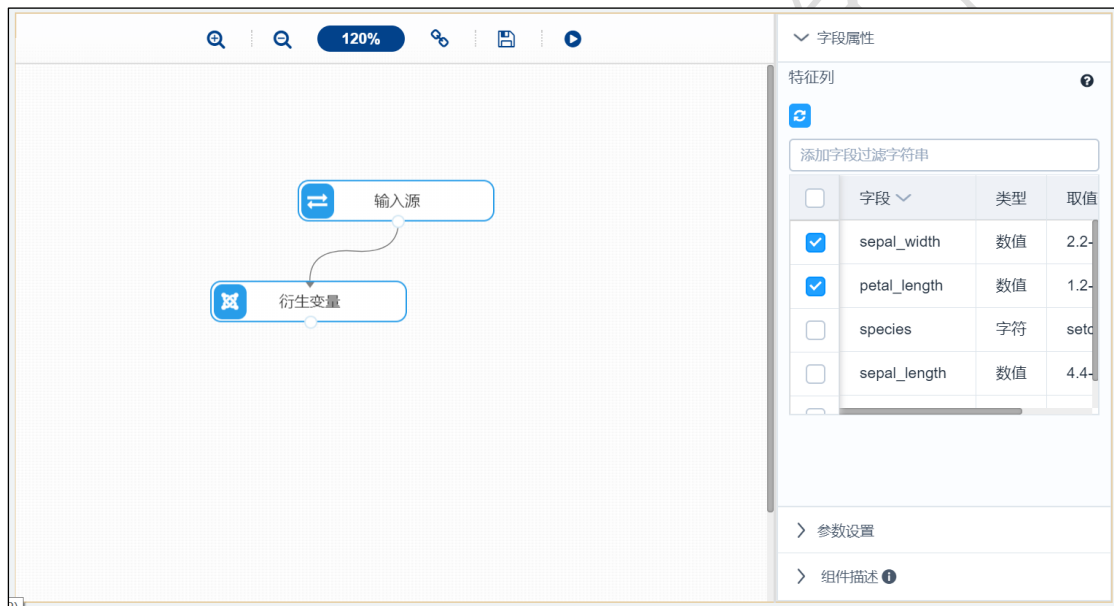


图 129



图 130

预览数据		
sepal_width	petal_length	new
3.5	1.4	4.9
3	1.4	4.4
3.2	1.3	4.5
3.1	1.5	4.6
3.6	1.4	5
3.9	1.7	5.6
3.4	1.4	4.8
3.4	1.5	4.9

共 150 条 25 条/页 < 1 2 3 4 5 6 > 前往 1 页

图 131

### 3.4.2.18 修改列名

图标:  修改列名

描述: 修改列名是指对数据表中的字段名进行修改。

字段属性

特征列: 必选。勾选列将传入下个组件,并且可供该组件修改字段名,如图 132 所示。

字段



添加字段过滤字符串

<input checked="" type="checkbox"/>	字段 ▾	类型	取值范围
<input checked="" type="checkbox"/>	id	数值	-
<input checked="" type="checkbox"/>	m	数值	1535.08-19...
<input checked="" type="checkbox"/>	r	数值	1-29
<input checked="" type="checkbox"/>	f	数值	1-24

图 132

列名转换：必填。修改对应的列名为需要的列名，如图 133 所示。

列名转换

原字段名	新字段名
id	id
m	mm
r	rr
f	ff

图 133

### 参数设置

无。

### 输出

表结果：列名修改后的结果。

报告：无。

### 示例

下面对某数据数据的列名进行修改。

- 勾选需要进行列名修改的数据，如果只勾选 m, r 两列，则只将 m, r 两列数据传

入下个组件。如图 134 所示。

- 运行该组件，右击选择查看数据，结果如图 135 所示。

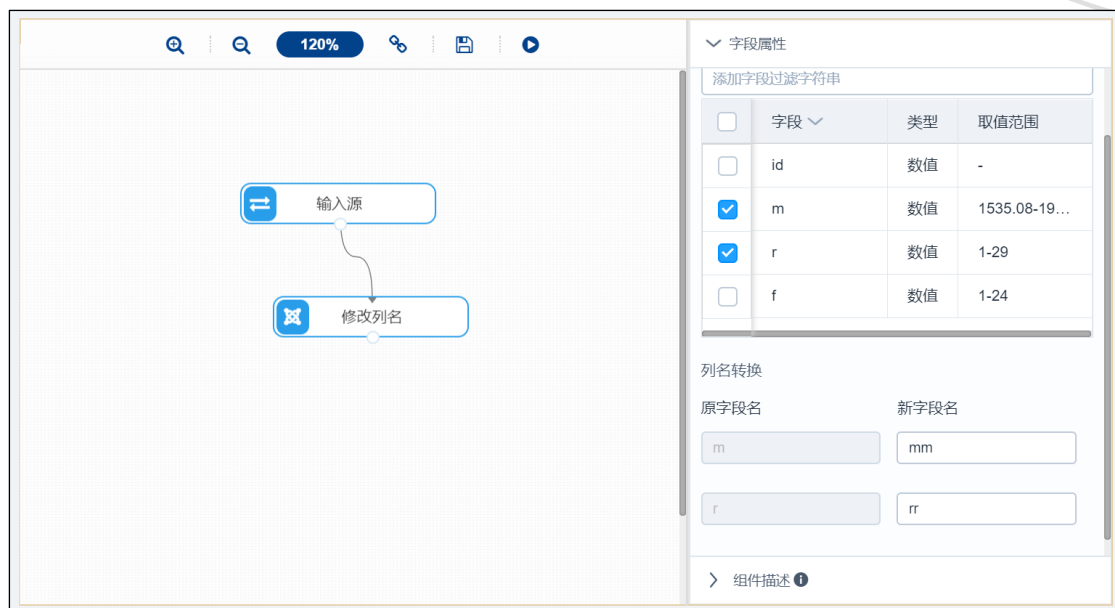


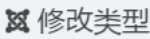
图 134

The '预览数据' (Preview Data) window displays a table with two columns: 'mm' and 'rr'. The data rows are as follows:

mm	rr
232.61	27
1507.11	3
817.62	4
232.81	3
1913.05	14
220.07	19
615.83	5
1059.66	26

图 135

### 3.4.2.19 修改类型

图标: 

**描述:** 修改类型是指对数据表中的字段类型进行修改, 目前平台提供的类型有数值型 (numeric)、字符型 (text)、日期 (date)、时间 (timestamp)。

**字段属性:**

字段: 必选。勾选需要修改类型的字段, 并传入下一个组件, 如图 136 所示。



图 136

**修改类型:** 当新类型为数值型时, 可以通过调整参数的值设定小数点位数, 如图 137 所示。



图 137

参数设置

无。

## 输出

表结果：修改类型之后的数据。

报告：无。

## 示例

下面对某数据的类型进行修改，原数据类型如图 138 所示。

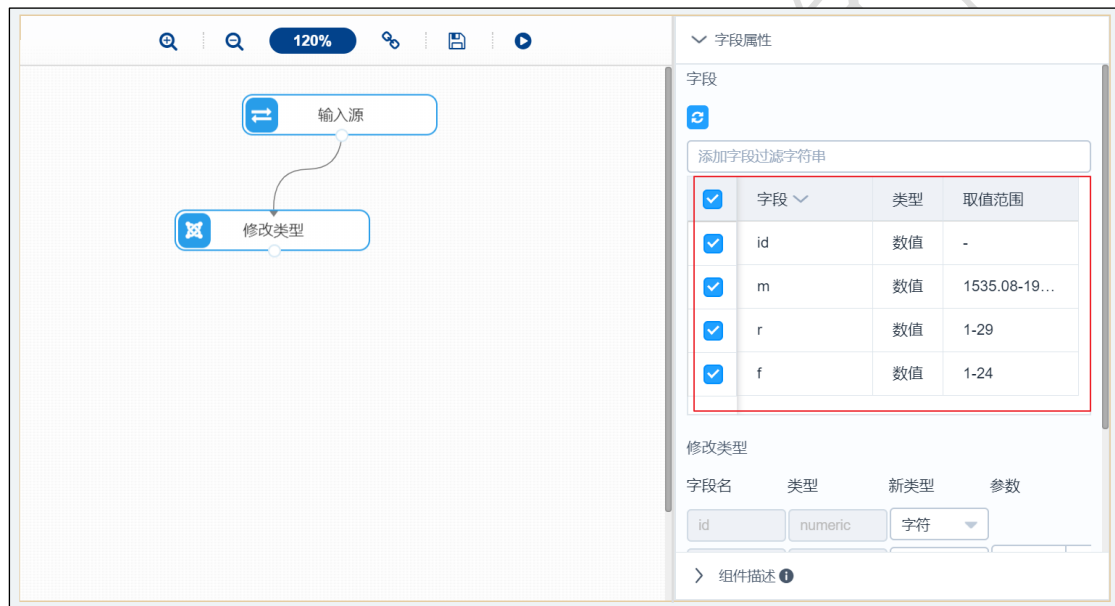


图 138

- 修改字段“id”的数据类型为字符型，如图 139 所示。
- 运行成功，下个组件就可获取到字段“id”的数据类型为字符型，如图 140 所示。



图 139

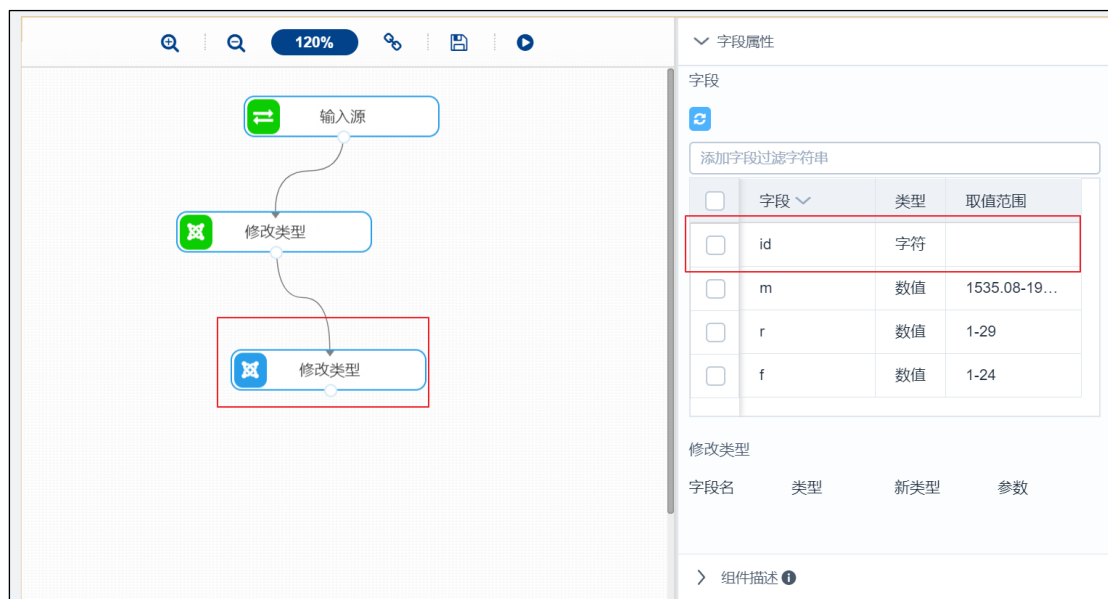


图 140

### 3.4.2.20 Python 脚本

图标: 

描述: Python 脚本是指可直接将 Python 脚本按照一定格式粘贴至脚本区作为组件运行。

字段属性

输入列表: 当输入节点连入某个数据之后, 则会将该数据的表名传值给 input1/input2/input3/input4, 如图 141 所示。

输入列表

input1 from 10005671\_1\_1

---

input2 from

---

input3 from

---

input4 from

---

图 141

脚本: 必填。在脚本区域输入 Python 脚本, 格式要求如以下示例。

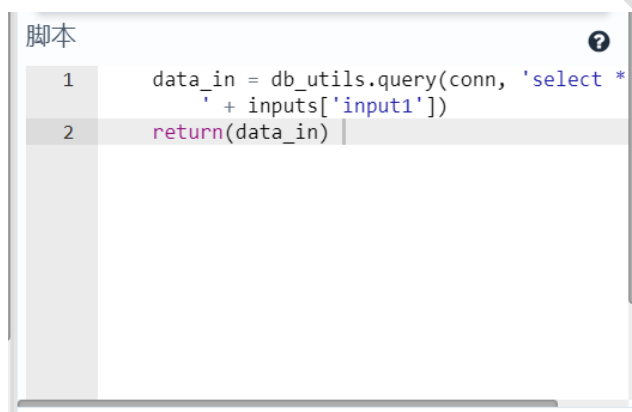


载入数据代码格式如以下两行脚本。

```
data_in = db_utils.query(conn, 'select '+ field1 + field2 +' from ' + inputs['input1'])  
return(data_in)
```

注意每行代码需要缩进 4 个空格

需要输出结构化数据，要求必须为 DataFrame，需要使用 return()。如图 142 所示



```
脚本  
1     data_in = db_utils.query(conn, 'select *  
    ' + inputs['input1'])  
2     return(data_in)
```

图 142

### 字段属性

无。

### 输出

表结果：由脚本运算出来的数据。

报告：无。

### 示例

对某数据执行 SQL 语句，原数据如图 143 所示。



sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa

共 150 条 25 条/页 < 1 2 3 4 5 6 > 前往 1 页

图 143

- 在脚本编辑区内放入 Python 代码，如图 144 所示。
- 运行成功后，查看数据，如图 145 所示。



图 144

sepal_length
5.1
4.9
4.7
4.6
5
5.4
4.6
5

图 145

### 3.4.2.21 Granger 因果检验



图标:

**描述:** Granger 因果检验是检验两列时序数据之间是否存在格兰杰因果关系。在时间序列情形下,两个经济变量 X、Y 之间的格兰杰因果关系定义为:若在包含了变量 X、Y 的过去信息的条件下,对变量 Y 的预测效果要优于只单独由 Y 的过去信息对 Y 进行的预测效果,即变量 X 有助于解释变量 Y 的将来变化,则认为变量 X 是引致变量 Y 的格兰杰原因。

### 字段属性

时间序列 1: 请选择数值型数据。

时间序列 2: 请选择数值型数据。

### 参数设置

延迟阶数: 默认 2。

### 输出

表结果: 无。

报告: 对应各滞后阶数的检验结果。

### 示例

下面对两个时序数据进行 Granger 因果检验。

- 选择两个时序数据,如图 146 所示。
- 设置延迟阶数为 2,则对每次 1-2 两个延迟阶数分别进行 Granger 因果检验。如图 147 所示。
- 运行成功,选择查看报告,如图 148 所示。



图 146

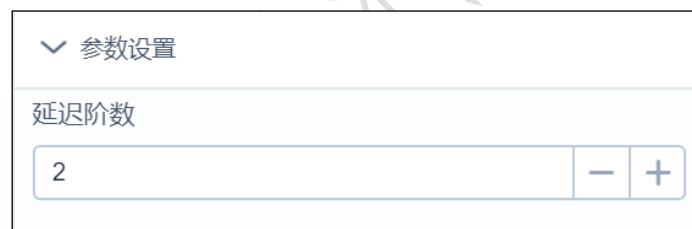


图 147



图 148