

# 数据挖掘-空值处理

## 概述

空值处理节点是将空值替换为均值、最大频数或者用户自定义的值等，实现空值的填充或者过滤。



- [概述](#)
- [输入/输出](#)
- [参数设置](#)



10.5.15版本后，左侧资源树不再显示空值处理节点，但是旧的空值处理节点仍能正常使用，建议使用最新的数据清洗节点完成空值替换操作。

## 输入/输出

|    |                     |
|----|---------------------|
| 输入 | 一个输入端口，用于接收数据集。     |
| 输出 | 一个输出端口，用于输出空值处理的结果。 |

## 参数设置

设置空值处理的参数：



参数面板用于设置需要处理空值的列，及空值处理的方式：单击 **选择列** 按钮，在弹出的窗口进行操作处理。

空值处理

分别用于关键字搜索可选和已选的字段

请输入关键字搜索

请输入关键字搜索

☐ 字段名称

☐ A产品大类名称

☐ #订单编号

☐ ACustomerID

☐ #年

☐ #月

☐ A区域

☐ AShipProvince

☐ A城市

用于勾选字段

☐ 字段名称

☐ A产品名称

☐ #数量

☐ #销售额

用于勾选字段

到右边>

<到左边

用于将勾选的字段选至左或右

全部

字符

数字

日期

用于按照数据类型筛选字段

| 字段名称  | 数据类型    | 处理策略          | 指定值 | 移除           |
|-------|---------|---------------|-----|--------------|
| A产品名称 | string  | 过滤整行          |     | <div>⊖</div> |
| #数量   | integer | 替换为最小值        |     | <div>⊖</div> |
| #销售额  | double  | 替换为最小值        | 0   | <div>⊖</div> |
|       |         | 替换为平均值        |     |              |
|       |         | 替换为中位数        |     |              |
|       |         | 删除空值占比高于百分比的列 |     |              |
|       |         | 替换为出现频率最高的值   |     |              |
|       |         | 替换为指定值        |     |              |
|       |         | 过滤整行          |     |              |

用于设置当前字段的空值处理策略

用于设置指定值或百分比

用于取消当前字段的空值处理策略

激活 Windows

转到“设置”以激活 Windows。

确定

取消

其中，处理策略包含如下内容：

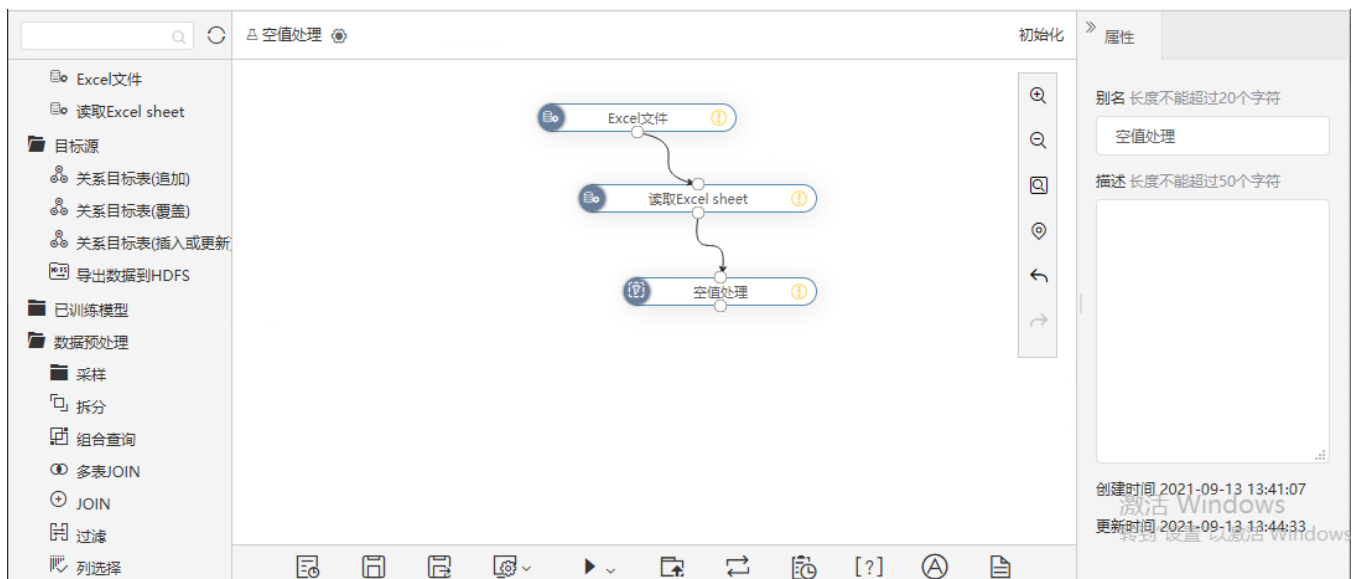
- 最大值：使用本列字段的最大值替换空值，适用于数值型字段。
- 最小值：使用本列字段的最小值替换空值，适用于数值型字段。
- 平均值：使用本列字段的平均值替换空值，适用于数值型字段。
- 中位数：使用本列字段的中位数替换空值，适用于数值型字段。
- 按空值百分比删除列：若空值超过设置的百分比，则删除整列；适用于字符、数值型字段。
- 出现频率最多替换：使用本列出现频率最多的值替换空值；适用于字符、数值型字段。
- 指定值：使用设定的值替换空值，适用于字符、数值型字段。此时，“指定值”列输入百分比，表示空值单元格占该列所有单元格达到指定百分比时，删除该列。
- 过滤整行：只要选定行有空值，就删除该行。此时，“指定值”列输入具体值，表示以该具体值来显示空值。

## 示例

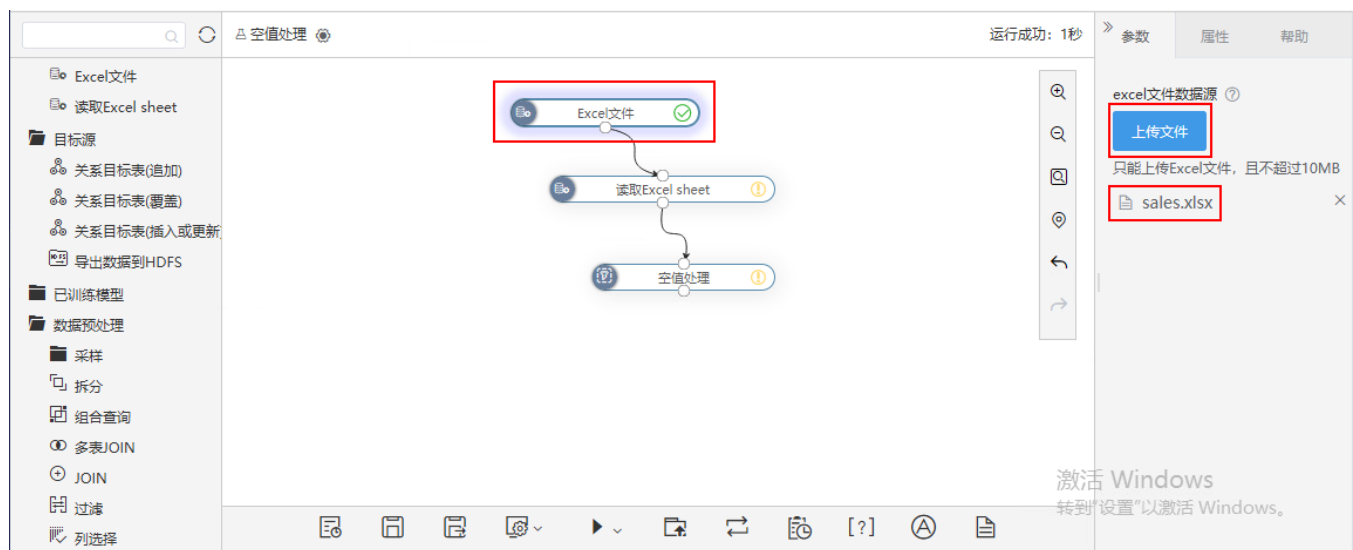
1、我们有如下Excel数据表：这张表中有几处空格。

|    | A      | B    | C     | D          | E    | F  | G            | H   | I  | J      |
|----|--------|------|-------|------------|------|----|--------------|-----|----|--------|
| 1  | 产品大类名称 | 产品名称 | 订单编号  | CustomerID | 年    | 区域 | ShipProvince | 城市  | 数量 | 销售额    |
| 2  | 特制品    | 沙茶   | 10249 | TOMSP      | 2016 | 华东 | 山东省          | 济南  | 9  | 167.4  |
| 3  | 海鲜     | 虾子   | 10250 | HANAR      | 2016 | 华北 | 河北省          | 秦皇岛 | 10 | 77     |
| 4  | 特制品    |      | 10250 | HANAR      | 2016 | 华北 | 河北省          | 秦皇岛 | 35 | 1261.4 |
| 5  | 调味品    |      | 10250 | HANAR      | 2016 | 华北 | 河北省          | 秦皇岛 | 15 | 214.2  |
| 6  | 谷类/麦片  | 小米   | 10251 | VICTE      | 2016 | 华东 | 江苏省          | 南京  | 15 | 222.3  |
| 7  | 调味品    | 海苔酱  | 10251 | VICTE      | 2016 |    | 江苏省          | 南京  |    |        |
| 8  | 点心     | 桂花糕  | 10252 | SUPRD      | 2016 | 东北 | 吉林省          | 长春  | 40 | 2462.4 |
| 9  | 饮料     | 汽水   | 10254 | CHOPS      | 2016 | 华中 | 湖北省          | 武汉  | 15 | 45.9   |
| 10 | 肉/家禽   | 鸭肉   | 10254 | CHOPS      | 2016 | 华中 | 湖北省          | 武汉  | 21 | 342.72 |
| 11 | 特制品    | 鸡精   | 10254 | CHOPS      | 2016 | 华中 | 湖北省          | 武汉  |    |        |
| 12 | 饮料     |      | 10255 | RICSU      | 2016 | 华北 | 北京市          | 北京  | 20 | 304    |
| 13 | 点心     | 饼干   | 10255 | RICSU      | 2016 | 华北 | 北京市          | 北京  | 35 | 486.5  |
| 14 |        |      |       |            |      |    |              |     |    |        |
| 15 |        |      |       |            |      |    |              |     |    |        |
| 16 |        |      |       |            |      |    |              |     |    |        |
| 17 |        |      |       |            |      |    |              |     |    |        |
| 18 |        |      |       |            |      |    |              |     |    |        |
| 19 |        |      |       |            |      |    |              |     |    |        |
| 20 |        |      |       |            |      |    |              |     |    |        |
| 21 |        |      |       |            |      |    |              |     |    |        |
| 22 |        |      |       |            |      |    |              |     |    |        |
| 23 |        |      |       |            |      |    |              |     |    |        |
| 24 |        |      |       |            |      |    |              |     |    |        |
| 25 |        |      |       |            |      |    |              |     |    |        |
| 26 |        |      |       |            |      |    |              |     |    |        |
| 27 |        |      |       |            |      |    |              |     |    |        |

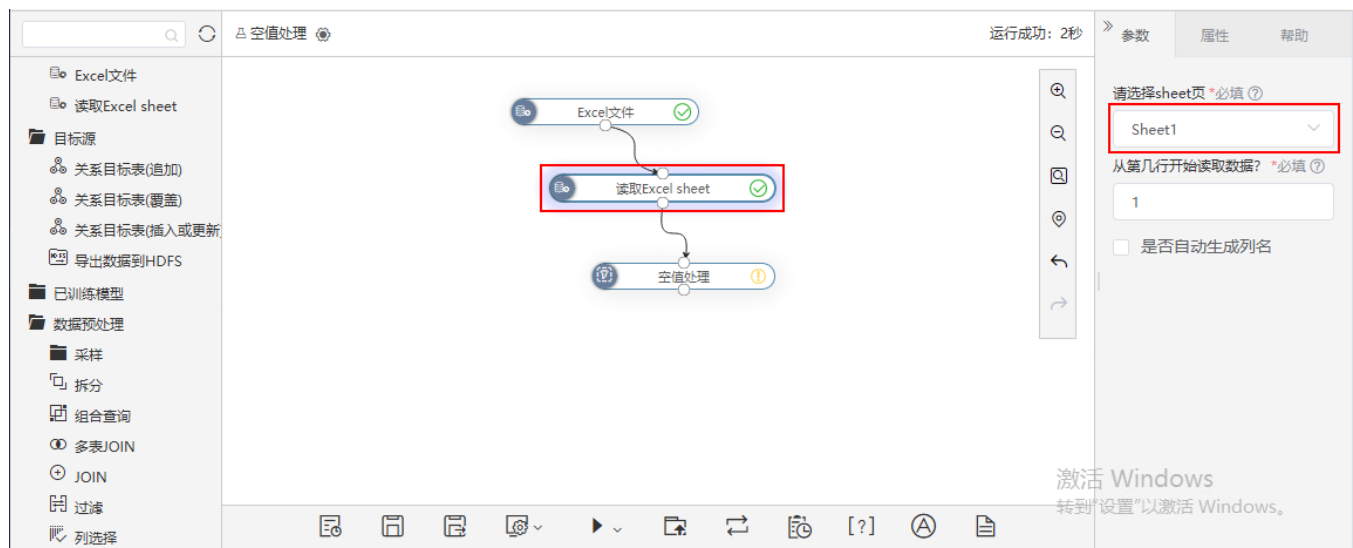
2、新建实验，在画布中拖拽如下节点：



3、在画布中选定“Excel文件”节点，参数面板中上传步骤1中的Excel数据表，并且执行该节点：



4、在画布中选定“读取Excel sheet”节点，参数面板中选择“sheet1”，并且执行该节点：



5、在画布中选定“空值处理”节点，参数面板中单击 **选择列**，弹出“空值处理”窗口，按照如下进行设置：



6、空值处理设置完成后，在画布中执行“空值处理”节点。执行完成后，我们通过查看输出浏览效果。下面显示处理前后的效果对比：

查看输出

×

当前显示 10 列 / 总共 10 列, 12 条 / 总共 12 条数据

列筛选 请选择

表头真名 表头别名

| 产品大类名称 | 产品名称 | 订单编号  | CustomerID | 年    | 区域   | ShipProvince | 城市  | 数量   | 销售额    |
|--------|------|-------|------------|------|------|--------------|-----|------|--------|
| 调味品    | 海苔酱  | 10251 | VICTE      | 2016 | null | 江苏省          | 南京  | null | null   |
| 点心     | 桂花糕  | 10252 | SUPRD      | 2016 | 东北   | 吉林省          | 长春  | 40   | 2462.4 |
| 饮料     | 汽水   | 10254 | CHOPS      | 2016 | 华中   | 湖北省          | 武汉  | 15   | 45.9   |
| 肉/家禽   | 鸭肉   | 10254 | CHOPS      | 2016 | 华中   | 湖北省          | 武汉  | 21   | 342.72 |
| 特制品    | 鸡精   | 10254 | CHOPS      | 2016 | 华中   | 湖北省          | 武汉  | null | null   |
| 饮料     | null | 10255 | RICSU      | 2016 | 华北   | 北京市          | 北京  | 20   | 304.0  |
| 点心     | 饼干   | 10255 | RICSU      | 2016 | 华北   | 北京市          | 北京  | 35   | 486.5  |
| 特制品    | 沙茶   | 10249 | TOMSP      | 2016 | 华东   | 山东省          | 济南  | 9    | 167.4  |
| 海鲜     | 虾子   | 10250 | HANAR      | 2016 | 华北   | 河北省          | 秦皇岛 | 10   | 77.0   |
| 特制品    | null | 10250 | HANAR      | 2016 | 华北   | 河北省          | 秦皇岛 | 35   | 1261.4 |
| 调味品    | null | 10250 | HANAR      | 2016 | 华北   | 河北省          | 秦皇岛 | 15   | 214.2  |
| 谷类/麦片  | 小米   | 10251 | VICTE      | 2016 | 华东   | 江苏省          | 南京  | 15   | 222.3  |

这三条数据被过滤掉了

查看输出

×

当前显示 10 列 / 总共 10 列, 9 条 / 总共 9 条数据

列筛选 请选择

表头真名 表头别名

| 产品大类名称 | 产品名称 | 订单编号  | CustomerID | 年    | 区域   | ShipProvince | 城市  | 数量 | 销售额    |
|--------|------|-------|------------|------|------|--------------|-----|----|--------|
| 调味品    | 海苔酱  | 10251 | VICTE      | 2016 | null | 江苏省          | 南京  | 9  | 0.0    |
| 点心     | 桂花糕  | 10252 | SUPRD      | 2016 | 东北   | 吉林省          | 长春  | 40 | 2462.4 |
| 饮料     | 汽水   | 10254 | CHOPS      | 2016 | 华中   | 湖北省          | 武汉  | 15 | 45.9   |
| 肉/家禽   | 鸭肉   | 10254 | CHOPS      | 2016 | 华中   | 湖北省          | 武汉  | 21 | 342.72 |
| 特制品    | 鸡精   | 10254 | CHOPS      | 2016 | 华中   | 湖北省          | 武汉  | 9  | 0.0    |
| 点心     | 饼干   | 10255 | RICSU      | 2016 | 华北   | 北京市          | 北京  | 35 | 486.5  |
| 特制品    | 沙茶   | 10249 | TOMSP      | 2016 | 华东   | 山东省          | 济南  | 9  | 167.4  |
| 海鲜     | 虾子   | 10250 | HANAR      | 2016 | 华北   | 河北省          | 秦皇岛 | 10 | 77.0   |
| 谷类/麦片  | 小米   | 10251 | VICTE      | 2016 | 华东   | 江苏省          | 南京  | 15 | 222.3  |

该处不处理



- 指定值不允许输入“null”，在Smartbi中空值默认显示为“null”；
- 不支持空值以空白单元格显示的效果。