第二章 知识发现过程与应用结构

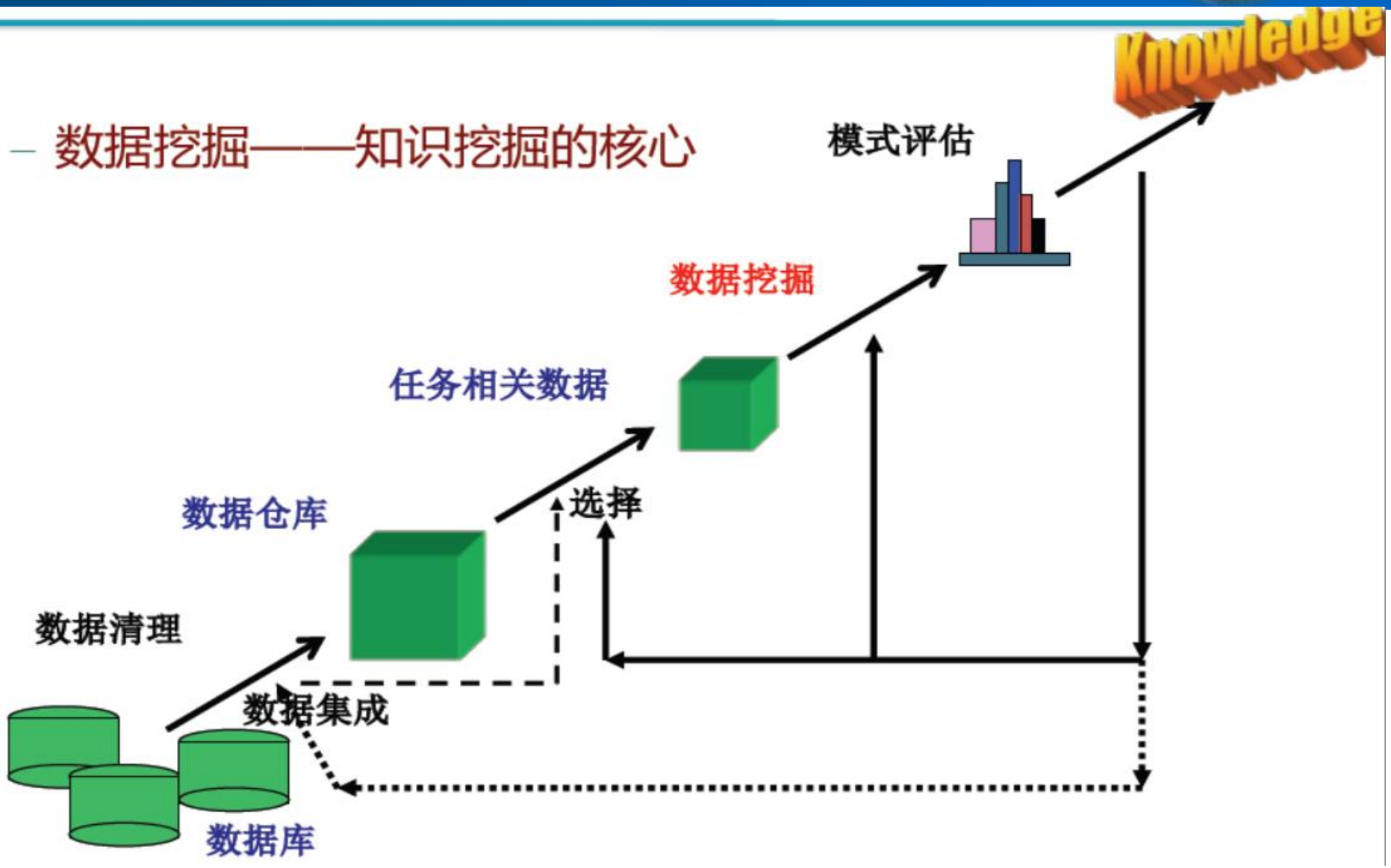
- 知识发现的基本过程
- 数据库中的知识发现处理过程模型
- 知识发现软件或工具的发展
- 知识发现项目的过程化管理
- 数据挖掘语言介绍

知识发现的基本过程

- 知识发现：Knowledge Discovery in Database，是一个系统化的工作，必须对可以利用的源数据进行分析，确定合适的挖掘目标，然后才能着手系统的设计和开发。
- 知识发现的过程可以简单地概括为：首先从数据源中抽取感兴趣的数据，并把它组织成适合挖掘的数据组织形式；然后，调用相应的算法生成所需的知识；最后对生成的知识模式进行评估，并把有价值的知识集成到企业的智能系统中。

- 一般地说，KDD是一个多步骤的处理过程，一般分为问题定义、数据采集、数据预处理、数据挖掘、模式评估等基本阶段。

知识发现的基本过程



– 数据挖掘——知识挖掘的核心

知识发现的基本过程

- KDD是一个多步骤的处理过程：
  - 1、问题定义
  - 2、数据采集
  - 3、数据预处理
  - 4、数据挖掘
  - 5、模式评估

数据挖掘实用案例分析——香水销售分析

任务：从某电商网站抓取1009条香水产品销售数据，包含香水产品的商品名称、产品毛重、商品产地、包装、香调、净含量、分类、适用性别、适用场所、价格和评价数

目标：分析香水销售的影响因素，为香水销售商判定采购计划以及用户选择香水提供依据。

数据挖掘实用案例分析——香水销售分析

从某电商网站上抓取到的香水产品销量数据，分析香水销售的影响因素，为香水销售商判定采购计划以及用户选择香水提供依据。

1、获取香水销售数据
2、香水销售数据预处理
3、香水销售数据统计分析
4、影响香水销量的因素分析
5、香水适用场所关联分析
6、香水聚类分析
7、香水营销建议

数据挖掘实用案例分析——香水销售分析

1、从某电商网站上抓取到的香水产品销量数据，分析香水销售的影响因素，为香水销售商判定采购计划以及用户选择香水提供依据。

2、香水销售数据预处理

3、香水销售数据统计分析

4、影响香水销量的因素分析

5、香水适用场所关联分析

6、香水聚类分析

7、香水营销建议

数据挖掘实用案例分析——香水销售分析

1	商品名称	商品产地	包装	香调	净含量	分类	性别	适用场所	价格	评价	旅行	其它	约会	情趣	商务	日常	party聚会	运动
2	冰希黎695600860	中国	Q版香水	花果香调	1ml-15ml	浓香水EDP	女	日常，约会，情趣，商务，party聚会，旅行	10	19000	1	0	1	1	1	1	1	0
3	冰希黎695600860	中国	Q版香水	混合香调	1ml-15ml	浓香水EDP	女	日常，约会，party聚会，运动，旅行	10	19000	1	0	1	0	0	1	1	1
4	（免邮）上海老		独立装	花果香调	31ml-100ml	浓香水EDP	女	日常，约会，商务，party聚会，旅行	18	90	1	0	1	0	1	1	1	0
5	法颂浪漫梦境女士		Q版香水	花果香调	1ml-15ml	固体香水/香膏	女	日常，约会，商务，party聚会，运动，旅行	22	30	1	0	1	0	1	1	1	1
6	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	日常，旅行	24	100	1	0	1	0	1	1	1	1
7	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	日常，约会，商务，旅行	24	100	1	0	0	0	0	0	1	0
8	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	日常，约会，商务，party聚会，运动，旅行	24	39000	1	0	0	0	0	1	0	0
9	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	party聚会，旅行	24	39000	1	0	1	0	1	1	0	0

香水销售数据预处理

Python编程处理 “评价”和“适用场所”字段：

1、“评价”字段的数据包含混合的中文和数字，末尾有一个“+”号，将其转为数值形式。即将类似“1.9万+”格式的“评价”字段的值转换为“19000”

2、“适用场所”分解为“旅行”、“其他”、“约会”、“工作”等8个字段，其类型是0和1,
将“商品产地”统一为“中国”

香水销售数据预处理

对香水产品的价格和评价数进行离散化处理，将价格和评价数离散化后的变量记为“价格等级”和“销量等级”：

1、将价格等间距地分为6个等级，记为低、较低、中等、较高、高、非常高

2、将评价数等间距地分为7个等级，记为非常低、低、较低、中等、较高、高、非常高

香水销售数据预处理

公式:

```
if (价格 <= 100) then '低'
 else if (价格 <= 300) then '较低'
  else if (价格 <= 500) then '中等'
   else if (价格 <= 700) then '较高'
    else if (价格 <= 1000) then '高'
     else '非常高'
    endif
   endif
  endif
 endif
endif
```

公式:

```
if (评价 <= 100) then '非常低'
 else if (评价 <= 500) then '低'
  else if (评价 <= 1000) then '较低'
   else if (评价 <= 2000) then '中等'
    else if (评价 <= 5000) then '较高'
     else if (评价 <= 10000) then '高'
      else '非常高'
     endif
    endif
   endif
  endif
 endif
endif
```

香水销售数据预处理

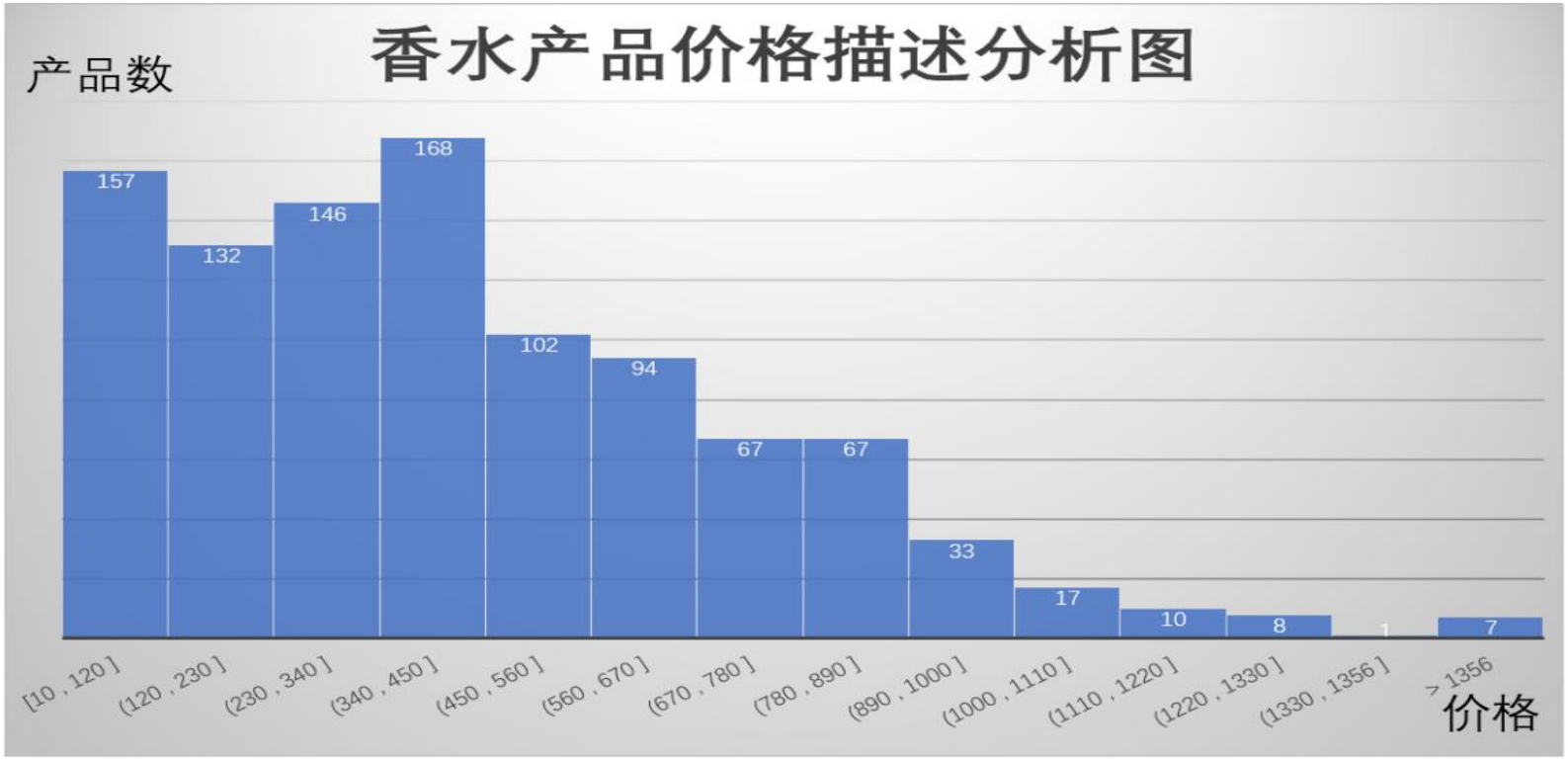
对香水产品的适用场合进行数量统计，得到新字段“适用场合数量”：

表格 注解

	商品名称	商品产地	包装	香调	净含里	分类	性别	适用场所	价格	评价	旅行	其它	约会	情趣	商务	日常	party...	运动	适用场合数里	价格等级	销里等级
1	冰希黎6...	中国	Q版香水	花果香调	1ml-15ml	浓香水EDP	女	日常，约会...	9.9...	190...	1.0...	0.0...	1.0...	1.0...	1.0...	1.0...	1.000	0.0...	6.000	低	非常高
2	冰希黎6...	中国	Q版香水	混合香调	1ml-15ml	浓香水EDP	女	日常，约会...	9.9...	190...	1.0...	0.0...	1.0...	0.0...	0.0...	1.0...	1.000	1.0...	5.000	低	非常高
3	（免邮...		独立装	花果香调	31ml-100ml	浓香水EDP	女	日常，约会...	18....	90.0...	1.0...	0.0...	1.0...	0.0...	1.0...	1.0...	1.000	0.0...	5.000	低	非常低
4	法颂浪...		Q版香水	花果香调	1ml-15ml	固体香水/...	女	日常，约会...	22....	30.0...	1.0...	0.0...	1.0...	0.0...	1.0...	1.0...	1.000	1.0...	6.000	低	非常低
5	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	日常，旅行	23....	100....	1.0...	0.0...	1.0...	0.0...	1.0...	1.0...	1.000	1.0...	6.000	低	非常低
6	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	日常，约会...	23....	100....	1.0...	0.0...	0.0...	0.0...	0.0...	0.0...	1.000	0.0...	2.000	低	非常低
7	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	日常，约会...	23....	390...	1.0...	0.0...	0.0...	0.0...	0.0...	1.0...	0.000	0.0...	2.000	低	非常高
8	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	party聚会...	23....	390...	1.0...	0.0...	1.0...	0.0...	1.0...	1.0...	0.000	0.0...	4.000	低	非常高
9	美顿香...		独立装	花果香调	31ml-100ml	淡香水EDT	女	日常，约会...	25....	50.0...	1.0...	0.0...	1.0...	0.0...	0.0...	1.0...	0.000	0.0...	3.000	低	非常低
10	冰希黎6...	中国	独立装	混合香调	其它	固体香水/...	女	日常，约会...	25....	190...	1.0...	0.0...	1.0...	0.0...	1.0...	1.0...	1.000	1.0...	6.000	低	非常高
11	冰希黎6...	中国	独立装	混合香调	其它	固体香水/...	女	日常，约会...	25....	190...	1.0...	0.0...	1.0...	1.0...	1.0...	1.0...	1.000	0.0...	6.000	低	非常高
12	艾诗止...		独立装	花果香调	31ml-100ml	香体走珠	女	日常，约会...	25....	300....	1.0...	0.0...	1.0...	0.0...	0.0...	1.0...	0.000	1.0...	4.000	低	低
13	艾诗止...		独立装	花果香调	31ml-100ml	香体走珠	女	日常，约会...	25....	300....	1.0...	0.0...	1.0...	0.0...	0.0...	1.0...	0.000	1.0...	4.000	低	低
14	阿迪达...	中国		海洋香调	31ml-100ml	香体走珠	女		26....	270...	0.0...	1.0...	0.0...	0.0...	0.0...	1.0...	0.000	1.0...	3.000	低	较高
15	阿迪达...	中国	独立装	混合香调	31ml-100ml	淡香水EDT	女	日常，运动...	26....	290...	1.0...	1.0...	1.0...	1.0...	1.0...	1.0...	1.000	1.0...	8.000	低	较高
16	阿迪达...	中国		混合香调	31ml-100ml	香体走珠	女	日常，约会...	26....	590...	0.0...	0.0...	0.0...	0.0...	0.0...	0.0...	0.000	0.0...	0.000	低	高
17	妮维雅...		独立装	花果香调	31ml-100ml	香体走珠	女	日常，约会...	28....	130...	0.0...	0.0...	1.0...	0.0...	0.0...	1.0...	1.000	1.0...	4.000	低	中等
18	Chanel...		独立装	花果香调	31ml-100ml	淡香水EDT	女	约会，情趣...	29....	40.0...	1.0...	0.0...	1.0...	0.0...	0.0...	1.0...	0.000	0.0...	3.000	低	非常低
19	美顿香...	中国	独立装	花果香调	31ml-100ml	淡香水EDT	女	日常，约会...	29....	300....	0.0...	0.0...	0.0...	0.0...	0.0...	1.0...	0.000	0.0...	1.000	低	低
20	美顿香...	中国	独立装	花果香调	31ml-100ml	淡香水EDT	女	日常，约会...	29....	300....	0.0...	0.0...	1.0...	1.0...	1.0...	0.0...	1.000	0.0...	4.000	低	低
21	美顿香...	中国	独立装	花果香调	31ml-100ml	淡香水EDT	女	日常，约会...	29....	300....	1.0...	0.0...	1.0...	0.0...	0.0...	1.0...	0.000	0.0...	3.000	低	低

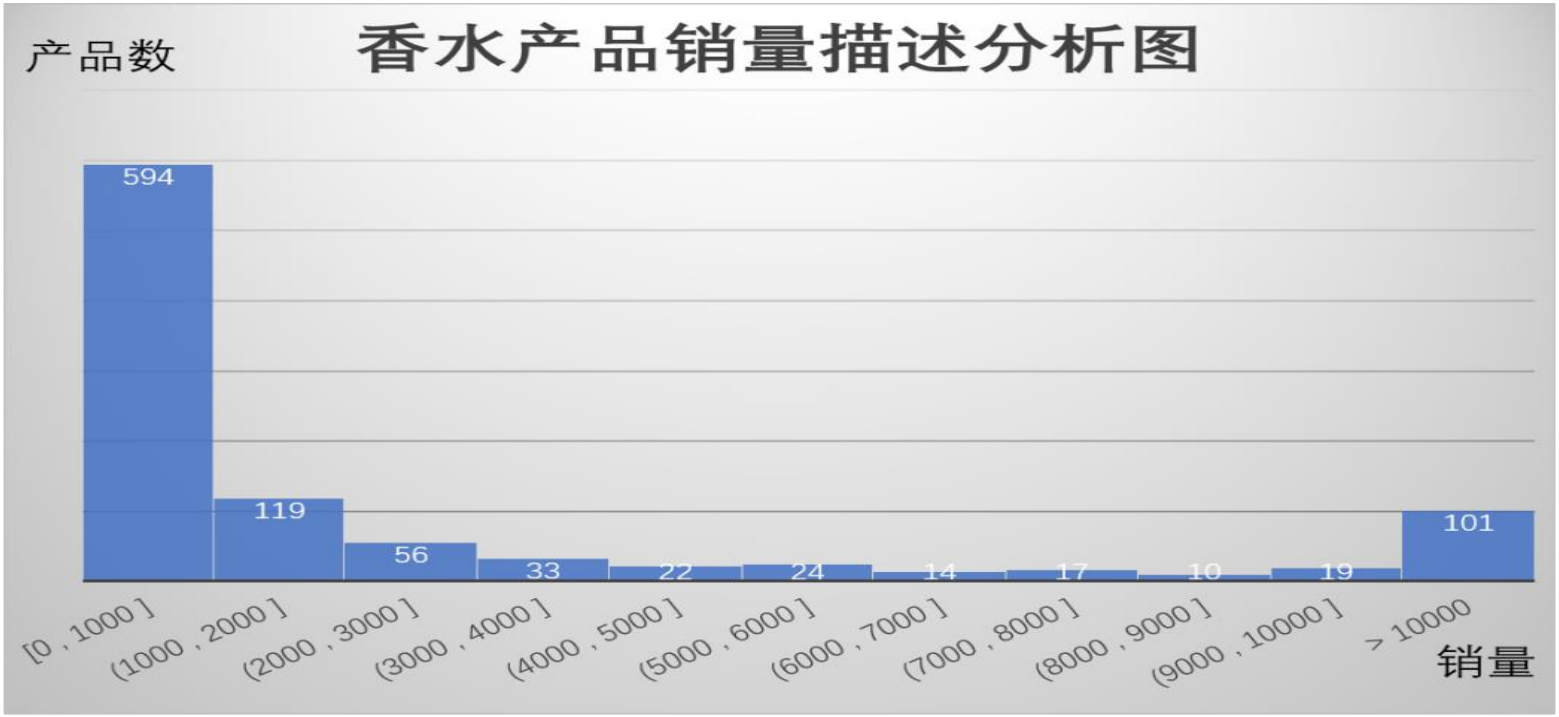
2、香水销售数据统计分析

香水产品价格描述分析图：



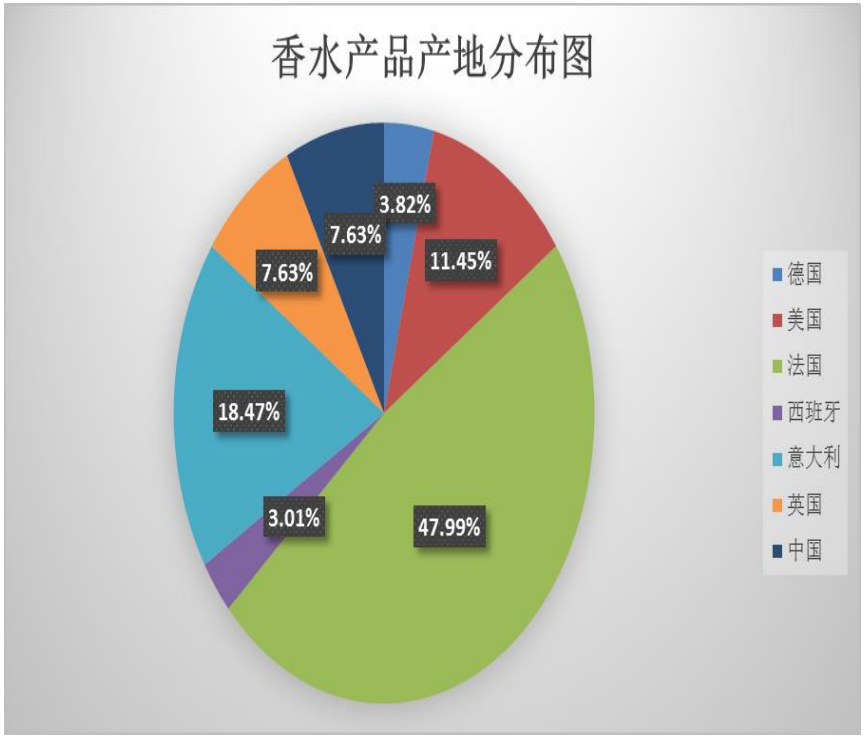
2、香水销售数据统计分析

香水产品销量描述分析图：

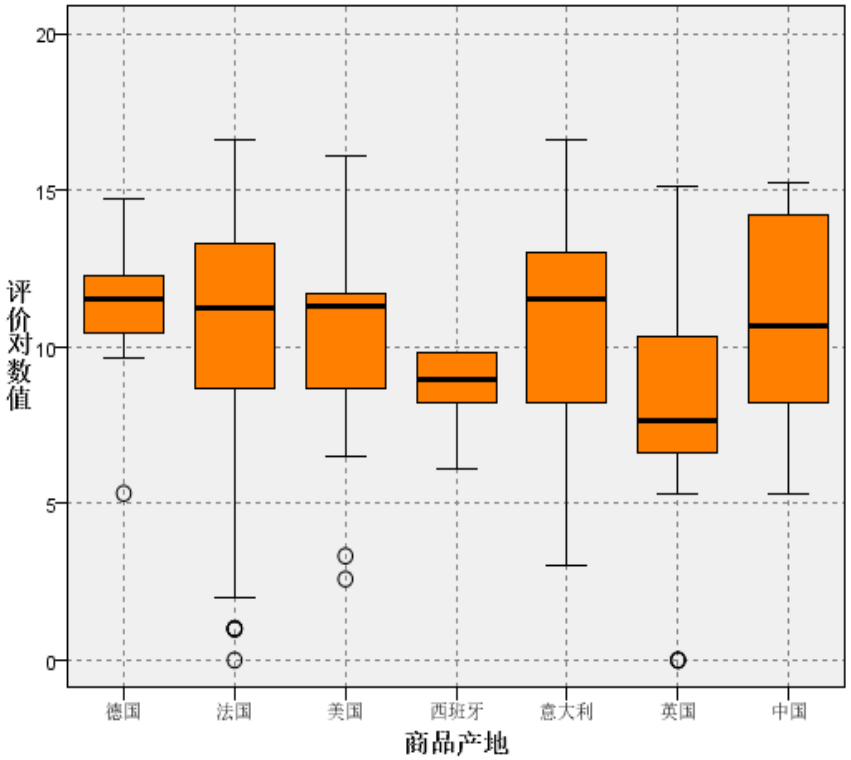


2、香水销售数据统计分析

香水产品产地分布图：

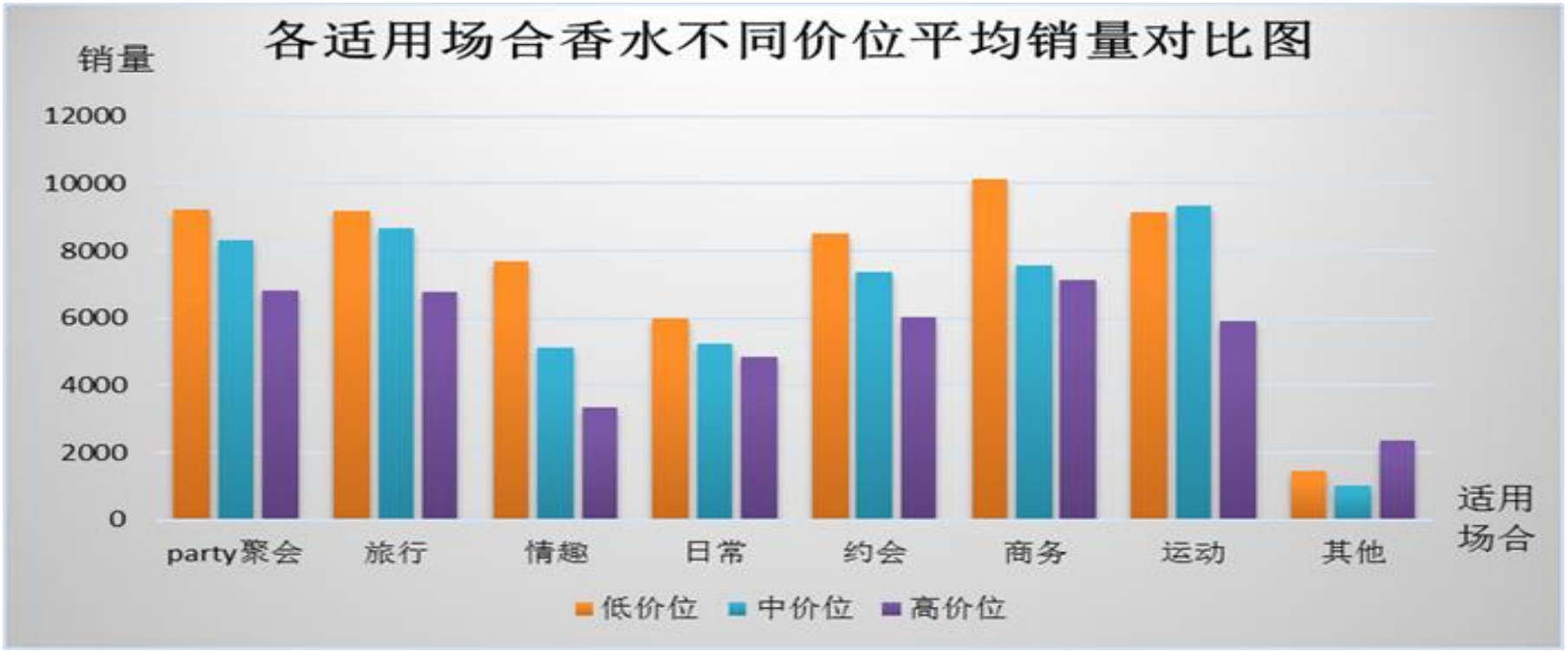


各产地香水销量箱型图



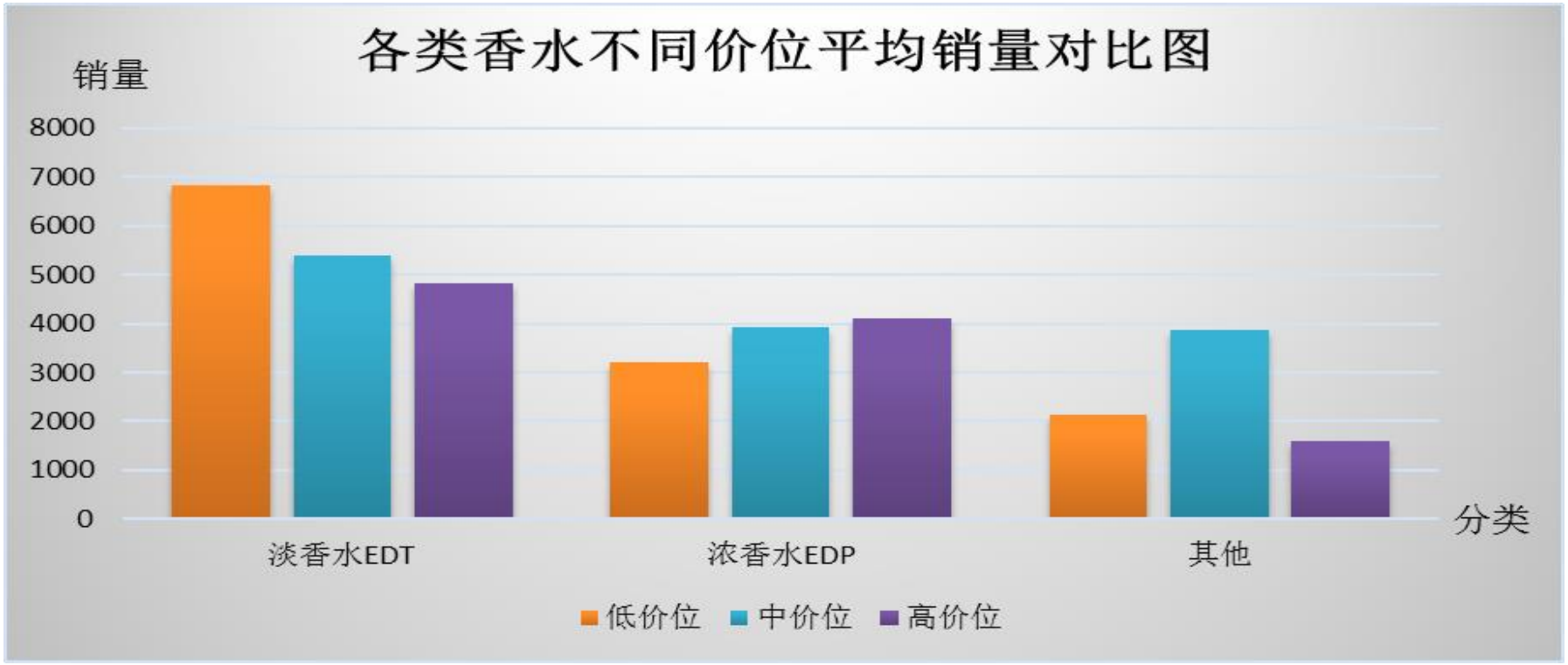
2、香水销售数据统计分析

各适用场合香水不同价位平均销量对比图



2、香水销售数据统计分析

各类香水不同价位平均销量对比图：



3、影响香水销量的因素分析

使用SPSS的“记录选项”-“选择”组件对“商品产地”和“包装”的空值数据进行过滤

3、影响香水销量的因素分析

使用“过滤器”节点，过滤掉本次分析不需要的字段

过滤器 注解

字段：已输入 21 个，已过滤 12 个，已重命名 0 个，已输出 9 个

字段	过滤器	字段
商品名称	✗	商品名称
商品产地	→	商品产地
包装	→	包装
香调	→	香调
净含量	→	净含量
分类	→	分类
性别	→	性别
适用场所	✗	适用场所
价格	✗	价格
评价	✗	评价
旅行	✗	旅行
其它	✗	其它
约会	✗	约会
情趣	✗	情趣
商务	✗	商务
日常	✗	日常
party聚会	✗	party聚会
运动	✗	运动
适用场合数量	→	适用场合数量
价格等级	→	价格等级
销量等级	→	销量等级

3、影响香水销量的因素分析

使用“类型”节点，将“销量等级”字段设置为目标，其他字段设置为输入

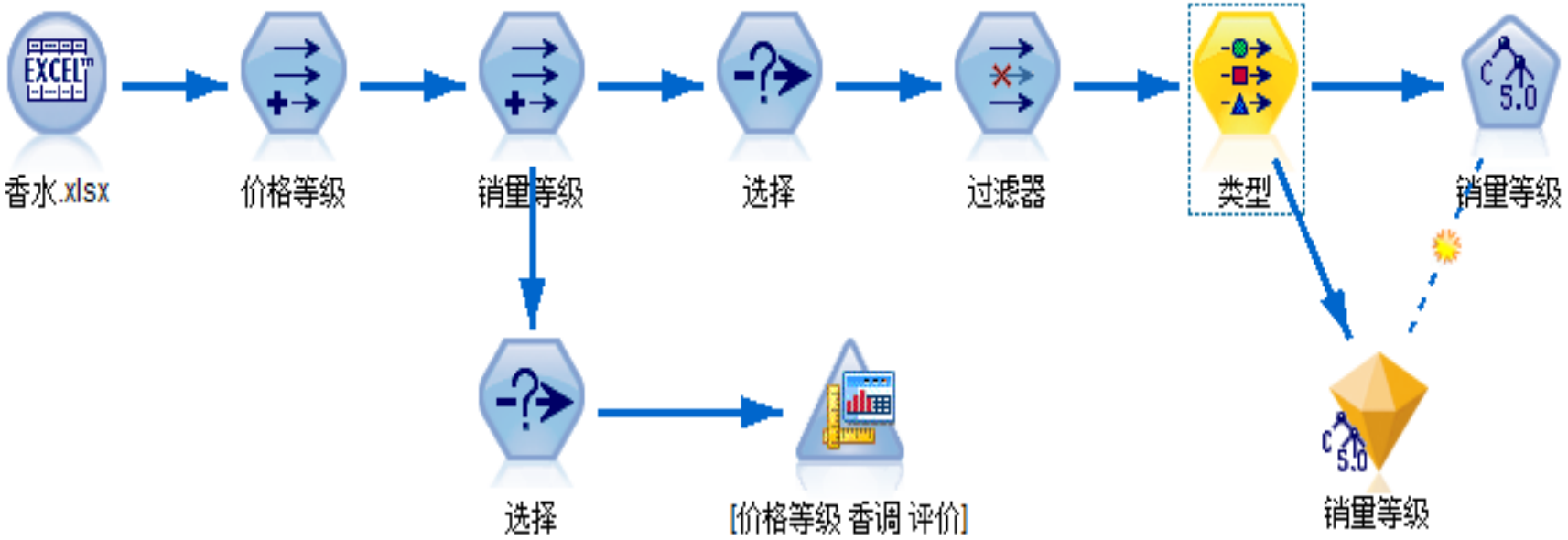
类型 格式 注解

读取值 清除值 清除所有值

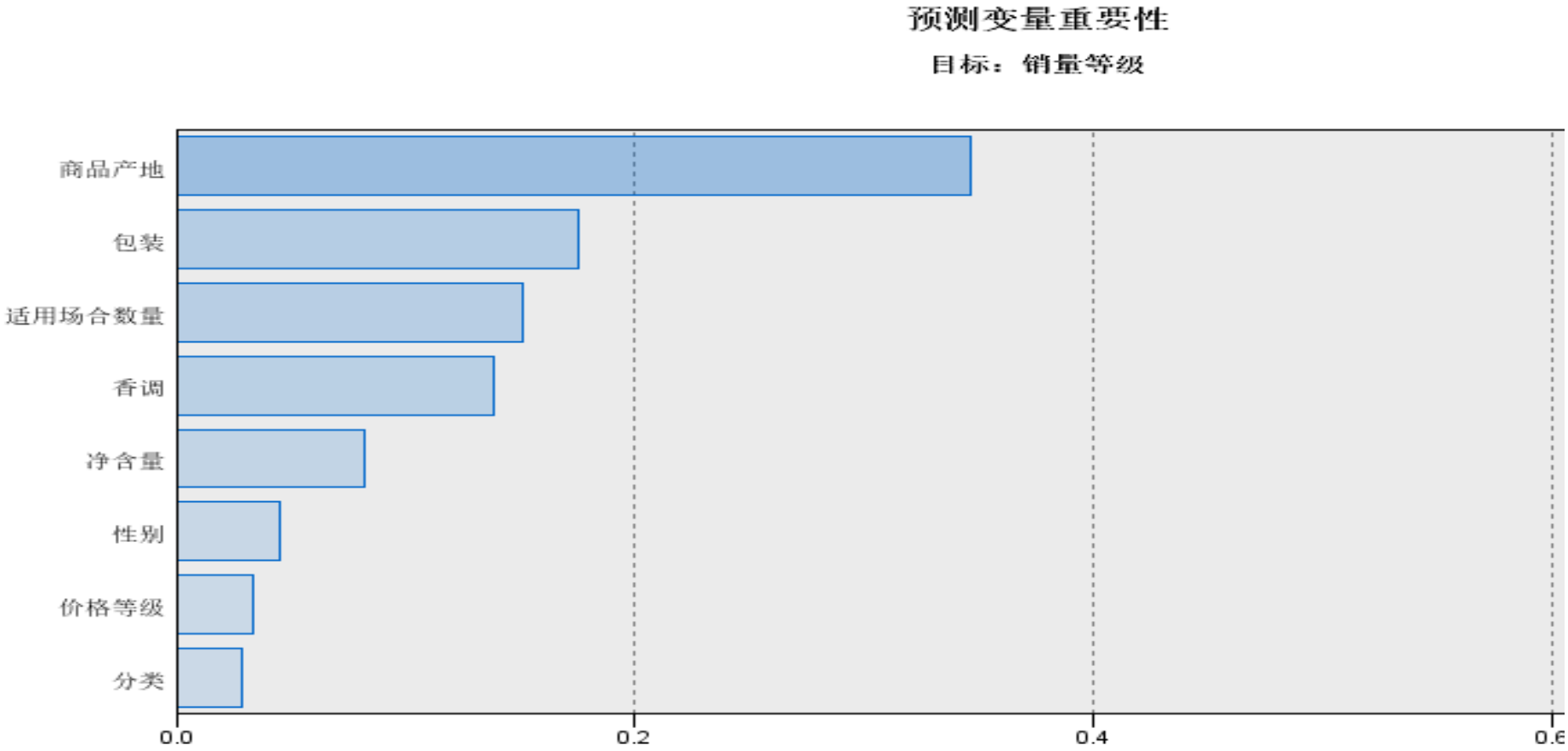
字段	测量	值	缺失	检查	角色
商品产地	名义	",中国,德...		无	输入
包装	名义	",Q版香水,...		无	输入
香调	名义	",东方香调...		无	输入
净含量	名义	","101ml-...		无	输入
分类	名义	",其它,发...		无	输入
性别	名义	女,通用		无	输入
适用场合数量	连续	[0.0,8.0]		无	输入
价格等级	名义	中等,低,较...		无	输入
销量等级	名义	中等,低,较...		无	目标

3、影响香水销量的因素分析

使用C4.5决策树算法，挖掘影响香水产品销量等级的因素

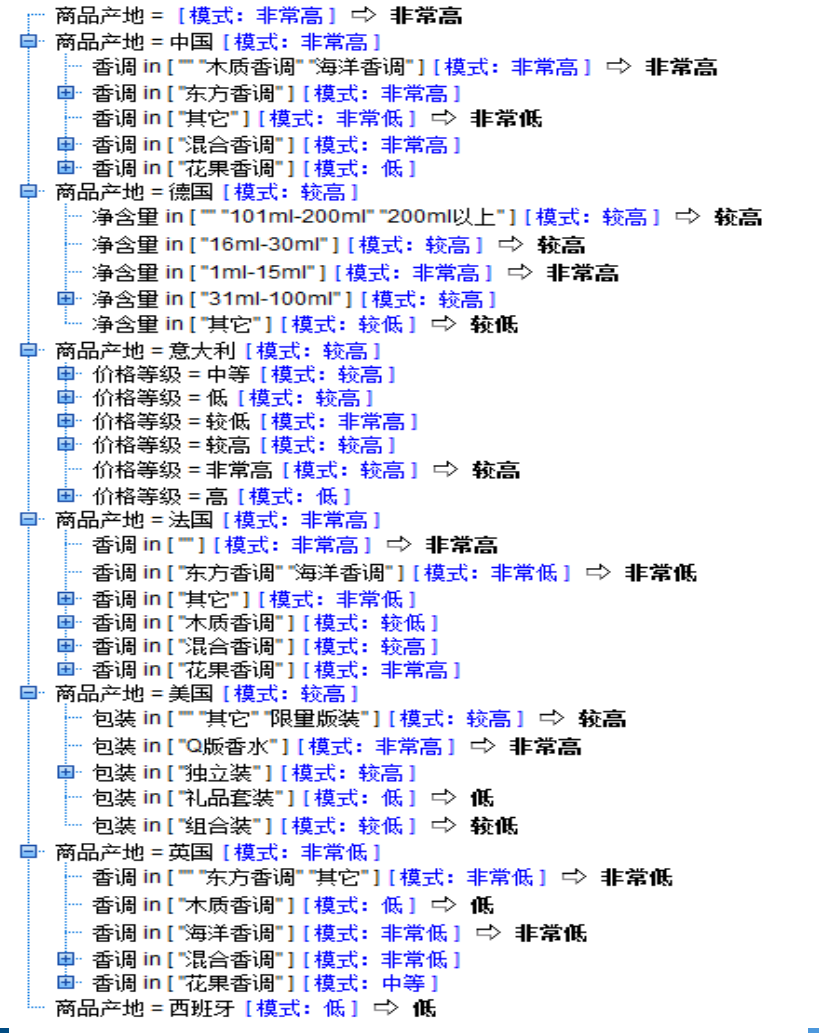


3、影响香水销量的因素分析



3、影响香水销量的因素分析

- 决策树
  - 中国和法国生产的香水最受消费者欢迎
  - 在中国和法国生产的香水中，消费者更加看重的是香水的香调。整体销量最高的“花果香调”在国产香水中销量反而较低，整体销量较低的“木质香料”销量却非常高
  - 德国、意大利和美国的香水整体销量较高；对德国香水，消费者更注重香水的净含量；对意大利香水，消费者更看重价格；对美国香水，消费者更看重包装
  - 英国和西班牙的产品销量较低。对英国香水，消费者更看重香调

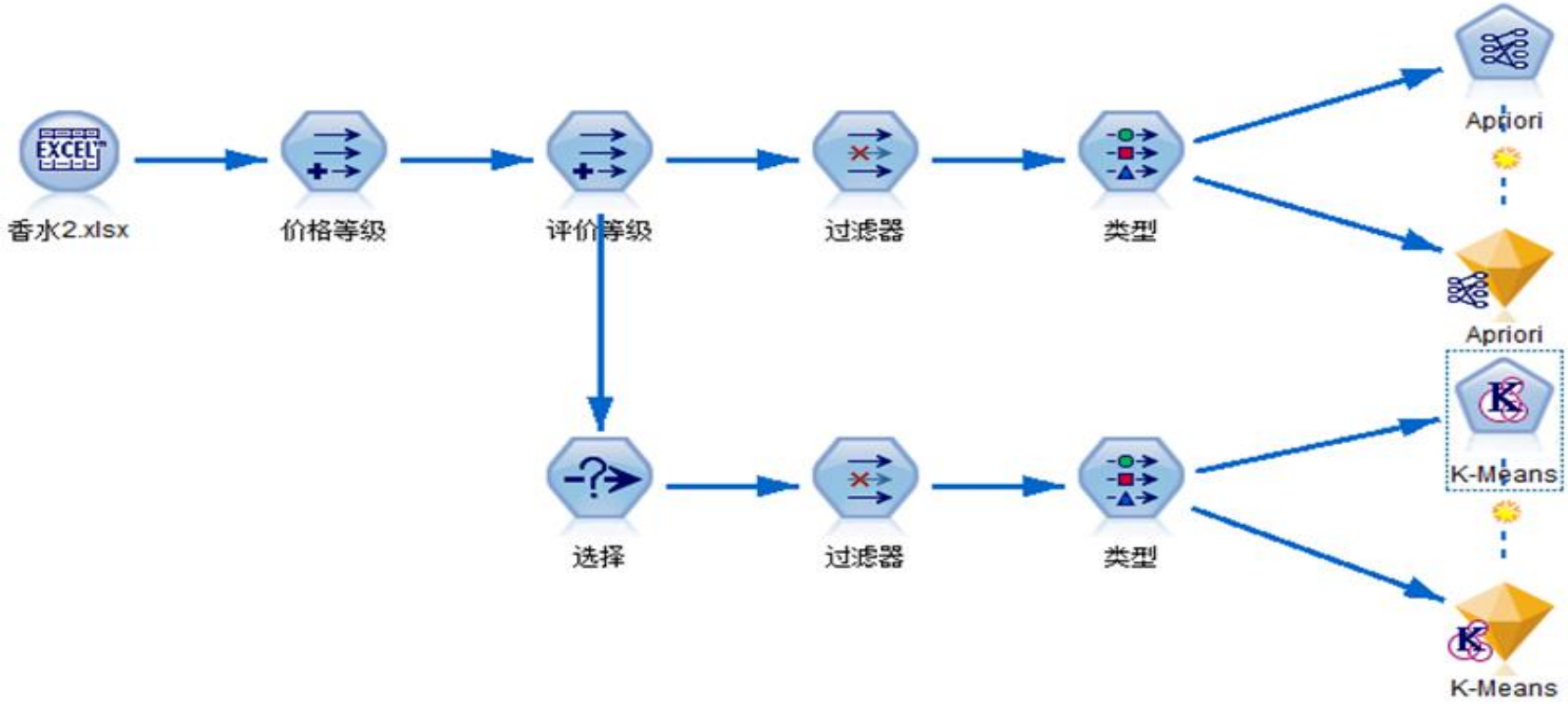


3、香水适用场合关联分析

对源数据进行预处理，将适用场所分隔开，生成不同的字段，总共为8类。含有该类适用场所的值设置为1.0，否则设置为0.0,在关联分析前滤除除适用场合外的所有本次分析不需要的字段，将所有适用场合的类型设置为任意

3、香水适用场合关联分析

采用Apriori算法，将最低条件支持度设为55%，最小规则置信度设置为90%



3、香水适用场合关联分析

大多数的适用场合之间关联性非常强

日常、商务、party聚会、约会出现次数最多，也是相互关联性最强的场所

排序依据: 置信度百分比 | 12 的 12

后项	前项	支持度百分比	置信度百分比
日常	商务	57.676	99.64
日常	party聚会 约会	56.432	98.897
日常	party聚会	60.373	98.625
约会	party聚会 日常	59.544	93.728
party聚会	商务	57.676	93.525
party聚会	商务 日常	57.469	93.502
约会	party聚会	60.373	93.471
日常	约会	66.805	93.168
约会	商务	57.676	91.727
约会	商务 日常	57.469	91.697
商务	party聚会 约会 日常	55.809	90.335
商务	party聚会 日常	59.544	90.244

4、香水聚类分析

将数据中的商品产地、包装、香调、净含量、分类、性别、适用场合数量作为输入字段进行聚类分析

使用过滤器节点过滤不需要的字段

过滤器 注解

字段：已输入 20 个，已过滤 11 个，已重命名 0 个，已输出 9 个

字段	过滤器	字段
商品名称	✗	商品名称
商品产地	→	商品产地
包装	→	包装
香调	→	香调
净含量	→	净含量
分类	→	分类
性别	→	性别
价格	✗	价格
评价	✗	评价
旅行	✗	旅行
其它	✗	其它
约会	✗	约会
情趣	✗	情趣
商务	✗	商务
日常	✗	日常
party聚会	✗	party聚会
运动	✗	运动
适用场合数量	→	适用场合数量
价格等级	→	价格等级
评价等级	→	评价等级

4、香水聚类分析

聚类类型节点设置

类型 格式 注解

读取值 清除值 清除所有值

字段	测量	值	缺失	检查	角色
商品产地	名义	中国,德国,意大利,法国,美国,英...		无	输入
包装	名义	Q版香水,独立装,礼品套装,组合...		无	输入
香调	名义	东方香调,其它,木质香调,海洋香...		无	输入
净含量	名义	"101ml-200ml","16ml-30ml","1...		无	输入
分类	名义	其它,发香雾,古龙水,固体香水/香...		无	输入
性别	标记	通用/女		无	输入
适用场合数量	名义	0.0,1.0,2.0,3.0,4.0,5.0,6.0,7.0,8.0		无	输入
价格等级	名义	中等,低,较低,较高,非常高,高		无	无
评价等级	名义	中等,低,较低,较高,非常低,非常...		无	无

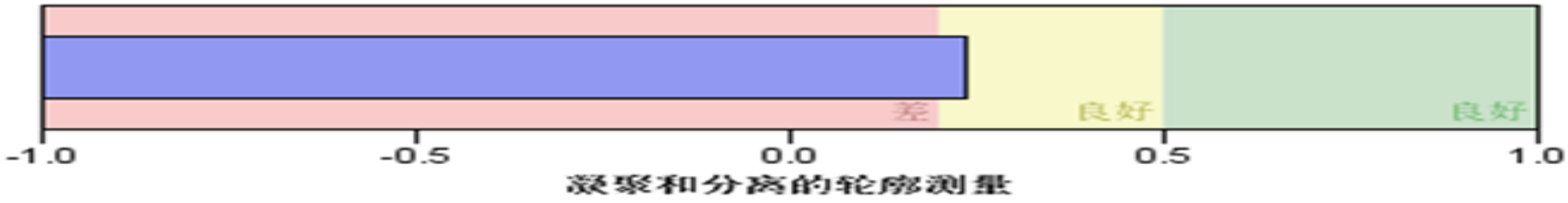
4、香水聚类分析

使用K-Means算法进行聚类，将聚类数设置为6得到聚类模型概要和聚类质量

模型概要

算法	K-Means
输入	7
聚类	6

聚类质量



4、香水聚类分析

K-Means聚类结果

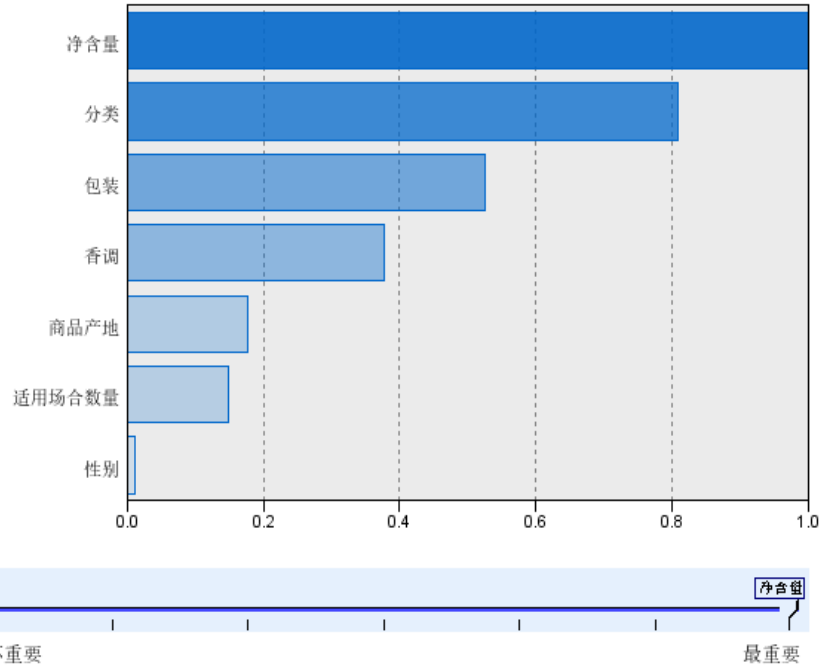
聚类

输入（预测变量）重要性

1.0 0.8 0.6 0.4 0.2 0.0

聚类	聚类-1	聚类-4	聚类-3	聚类-6	聚类-2	聚类-5
标签						
描述						
大小	24.5% (112)	23.0% (105)	22.1% (101)	11.6% (53)	9.6% (44)	9.2% (42)
输入	净含量 31ml-100ml (66.1%)	净含量 31ml-100ml (98.1%)	净含量 31ml-100ml (74.3%)	净含量 31ml-100ml (79.2%)	净含量 1ml-15ml (90.9%)	净含量 16ml-30ml (97.6%)
	分类 淡香水EDT (100.0%)	分类 浓香水EDP (100.0%)	分类 淡香水EDT (100.0%)	分类 淡香水EDT (98.1%)	分类 淡香水EDT (79.5%)	分类 浓香水EDP (95.2%)
	包装 独立装 (94.6%)	包装 独立装 (96.2%)	包装 独立装 (94.1%)	包装 独立装 (92.5%)	包装 Q版香水 (61.4%)	包装 独立装 (92.9%)
	香调 花果香调 (98.2%)	香调 花果香调 (62.9%)	香调 花果香调 (96.0%)	香调 混合香调 (73.6%)	香调 花果香调 (52.3%)	香调 花果香调 (73.8%)
	商品产地 意大利 (41.1%)	商品产地 法国 (61.0%)	商品产地 法国 (75.2%)	商品产地 法国 (50.9%)	商品产地 法国 (29.5%)	商品产地 法国 (57.1%)
	适用场合数量 1.000 (52.7%)	适用场合数量 1.000 (22.9%)	适用场合数量 6.000 (57.4%)	适用场合数量 1.000 (30.2%)	适用场合数量 1.000 (29.5%)	适用场合数量 1.000 (33.3%)
	性别 女 (96.4%)	性别 女 (99.0%)	性别 女 (94.1%)	性别 女 (92.5%)	性别 女 (90.9%)	性别 女 (100.0%)

预测变量重要性



5、香水营销建议

制定价格方面：将产品价格定位在大众消费品的水平，并保持正常利润空间。对淡香水EDT类产品，在不亏损的前提下适当降低价格；对浓香水EDT类产品，在调整空间内适当提高价格

产品分类方面：香水产品的产地、香调、净含量等对销售有很大影响，需要综合考虑几种因素才能获得更多收入

销售策略方面：消费者在购买香水产品时体现出了明显的价格敏感性，价格低的销量更好。组合装的香水销量好于其他包装。商家需要结合不同使用场合推出更多的香水组合和礼品装香水，以刺激消费

数据挖掘实用案例分析——香水销售分析

从某电商网站上抓取到的香水产品销量数据，分析香水销售的影响因素，为香水销售商判定采购计划以及用户选择香水提供依据。

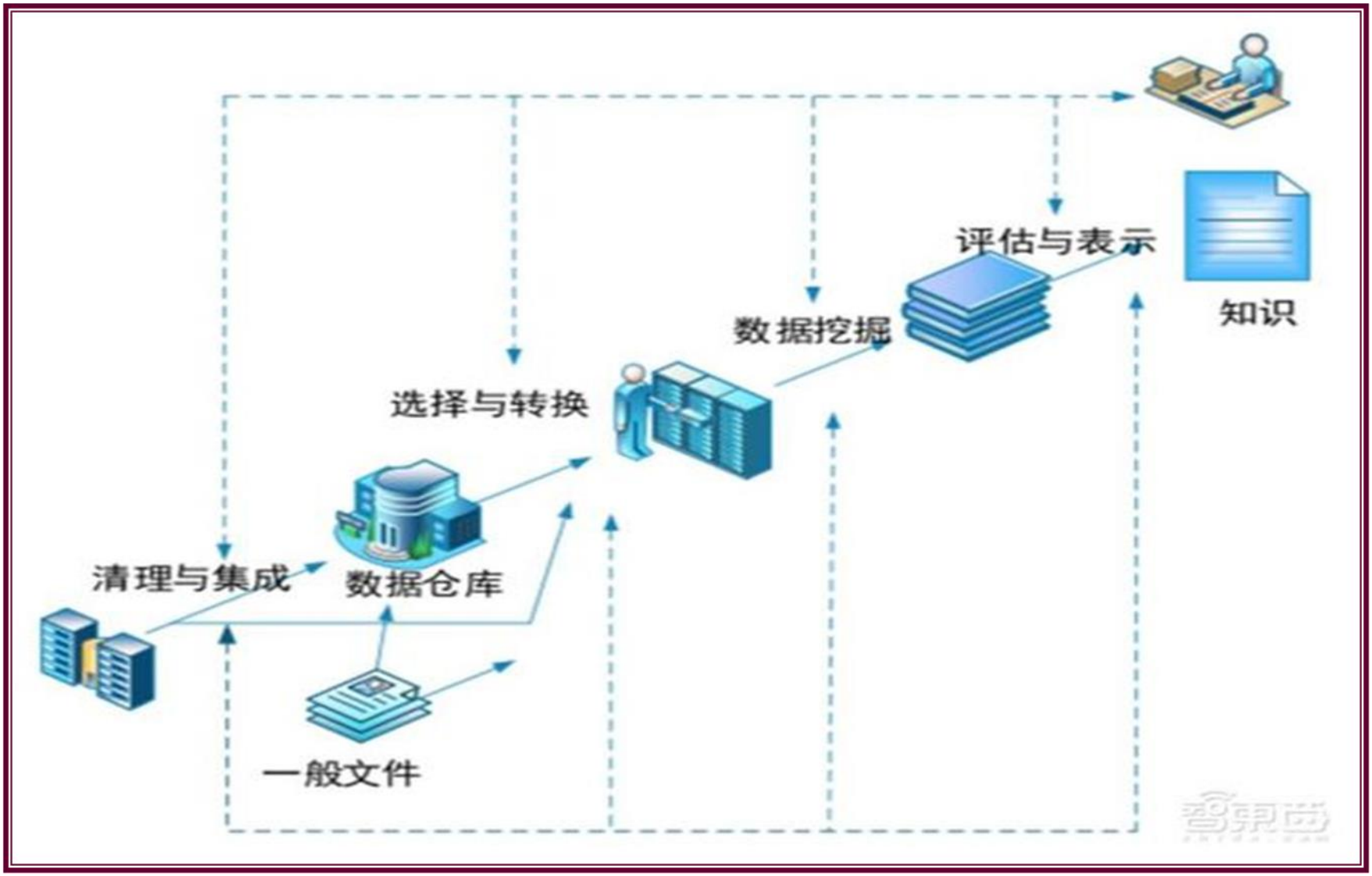
1、获取香水销售的相关数据
2、香水销售数据预处理
3、香水销售数据统计分析
4、影响香水销量的因素分析
5、香水适用场所关联分析
6、香水聚类分析
7、香水营销建议

数据挖掘的过程

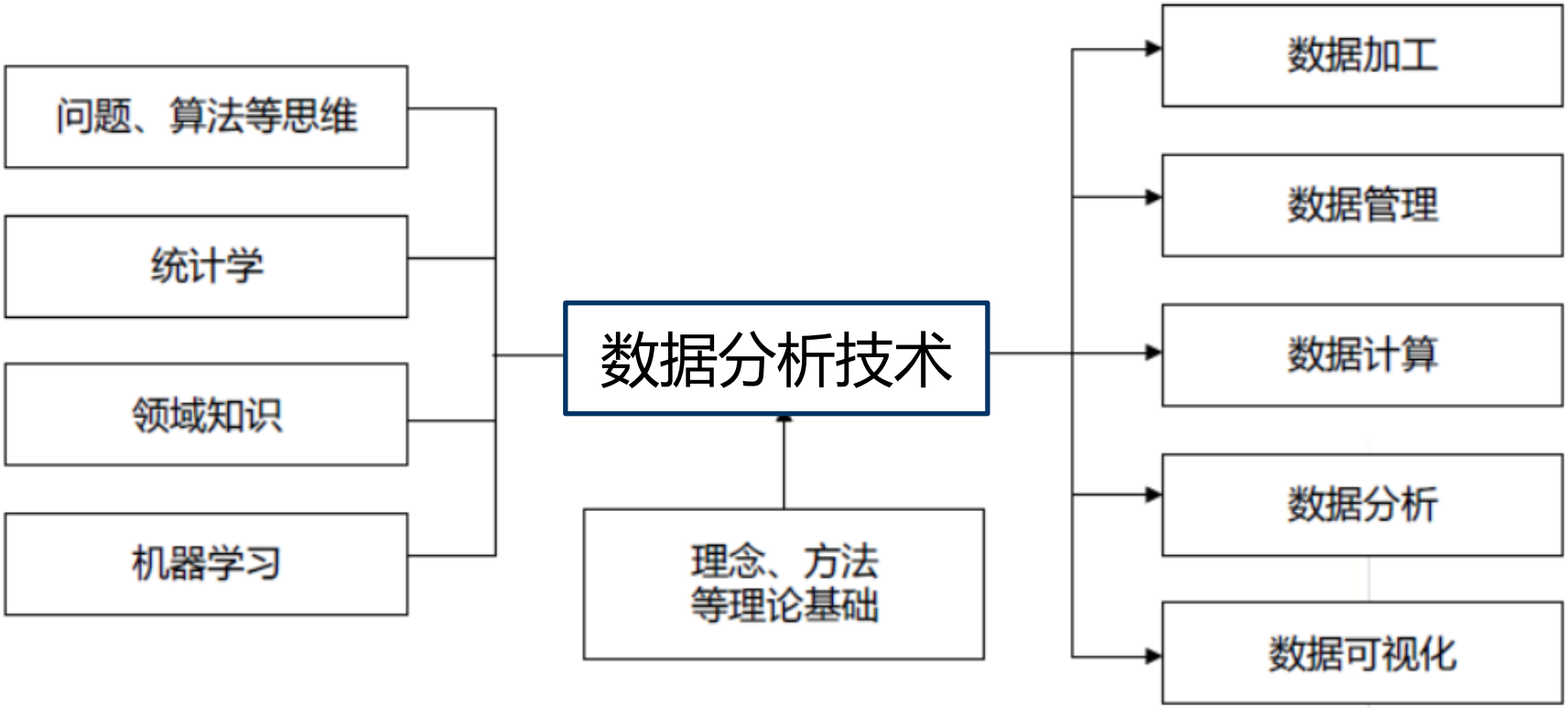
- 数据（Data）、信息（Information）和知识（Knowledge）是广义数据表现的不同形式。

data → information ↔ knowledge

数据挖掘的一般过程



交叉学科


问题、算法等思维
统计学
领域知识
机器学习
数据分析技术
理念、方法
等理论基础
数据加工
数据管理
数据计算
数据分析
数据可视化