****

**香水销售案例报告**

学 院： 金融学院

课程名称： 数据挖掘

教学班级： A01

姓 名： 李茂源

学 号： 0172841

目录

[1 引言 3](#_Toc9193949)

[1.1 案例背景 3](#_Toc9193950)

[1.2 任务理解 3](#_Toc9193951)

[2 数据理解 3](#_Toc9193952)

[2.1 原始数据表的理解 3](#_Toc9193953)

[2.2 数据审核 3](#_Toc9193954)

[2.3 数据分布 4](#_Toc9193955)

[2.3.1 价格 4](#_Toc9193956)

[2.3.2 评价 4](#_Toc9193957)

[2.3.3 产地 5](#_Toc9193958)

[3 数据预处理 6](#_Toc9193959)

[3.1 相关性分析 6](#_Toc9193960)

[3.1.1 商品产地和价格 6](#_Toc9193961)

[3.1.2 性别和分类 7](#_Toc9193962)

[3.1.3 毛重和净含量的关系 7](#_Toc9193963)

[3.1.4 分类和评价的关系 8](#_Toc9193964)

[3.1.5 商品产地和评价的关系 9](#_Toc9193965)

[3.1.6 包装和评价的关系 10](#_Toc9193966)

[3.1.7 适用场所的相关性分析 10](#_Toc9193967)

[3.1.8 适用场所、香调和分类之间的关系 11](#_Toc9193968)

[3.1.9 适用场所数量 12](#_Toc9193969)

[3.2 缺失值填充 12](#_Toc9193970)

[3.3 离散化数值 13](#_Toc9193971)

[3.3.1 价格离散化 13](#_Toc9193972)

[3.3.2 评价的离散化 13](#_Toc9193973)

[4 建立模型 15](#_Toc9193974)

[5 建议和意见 16](#_Toc9193975)

# 引言

## 案例背景

近年来，我国香水市场快速发展，同时市场内也涌现了大量的不同香水产品。那么到底什么样的香水产品销量更好，更受消费者欢迎呢？这些香水产品又有哪些特点呢？通过探索香水销售数据，来进行定向销售香水。

## 任务理解

对一个香水销售数据进行数据挖掘，了解消费者的购买习惯，使用场景，价格之间的关系，做到销售更加有效率。

# 数据理解

## 原始数据表的理解

香水销售数据表在名称中有很多的非常相似的记录，很多时候仅仅是适用场景不同，是不是需要合并。相同香水名称的评价都是一样的，评价应该就是对商品的评价，商品的评价有可能是和销量是正相关。产品的毛重带有单位，而且单位不统一，可能需要后续处理。

## 数据审核

整个数据有1009条记录，18个字段。其中商品产地的缺失值较多，有520条缺失，超过一半的缺失值，考虑填充。其中使用场合默认是连续型，但是应该改成标称型的。价格和评价都有离群值，查看后发现价格最大为2212，应该是在可接受范围内，且只有一例，要特殊考虑，不能删除。

## 数据分布

仅仅通过数据审核节点不能够很好掌握数据分布，需要用各种图来进行分析。

### 价格

通过数据审核我们知道价格跨度非常大，最小仅仅9.9，最大有2210。我们通过直方图来观察一下。

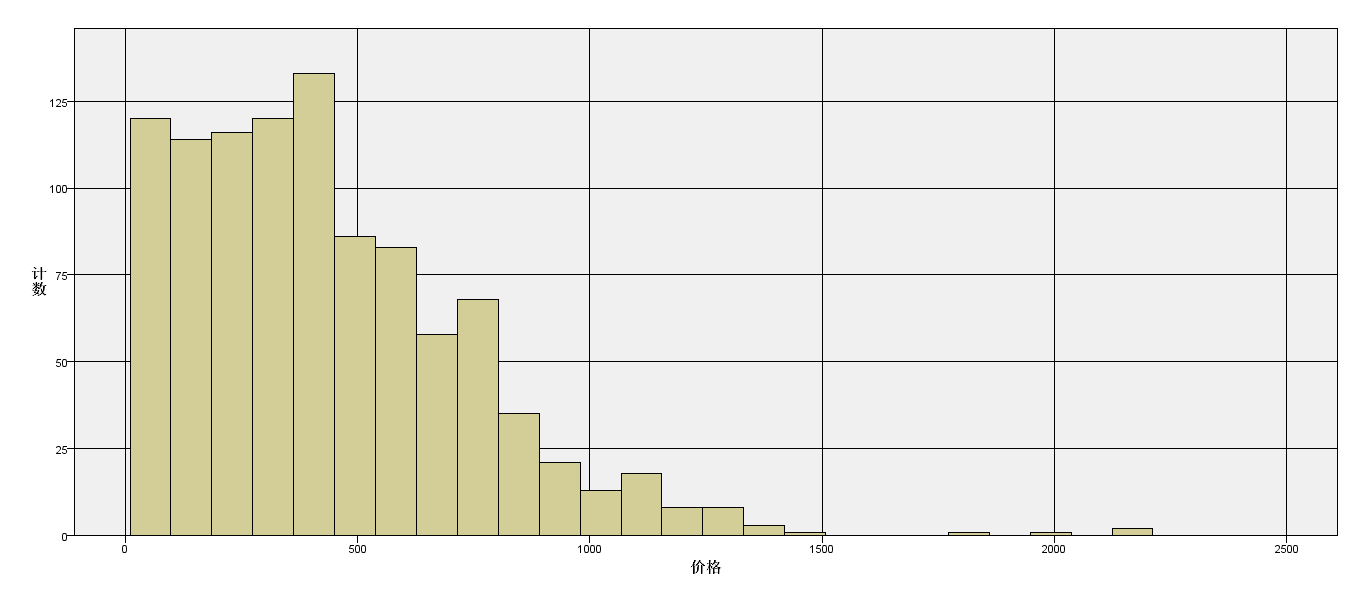


图 2‑1

从图2-1可以看出：

* 价虽然跨度很大，但是大部分分布在1000以下，说明1000以上的价格都算挺高的了。
* 等会可以考虑对价格进行离散化。
* 价格最高的2210为香奈儿机会/机遇/黄色邂逅女士香水50/100ML/持久淡香精EDP EDP100ml。
* 因为极差实在太大，在作图时候就要使用对数进行处理，这样可以方便观察。

### 评价

因为评价与销量是正相关，通过观察评价分布可以得到香水销量分布。

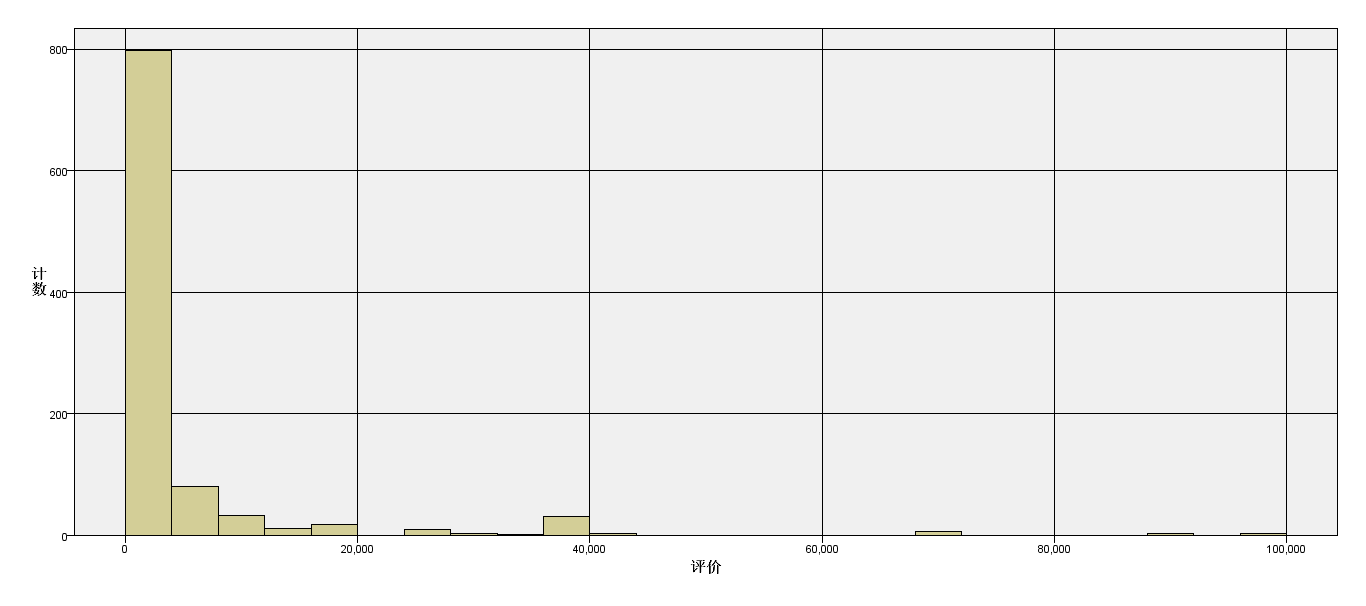


图 2‑2

如图2-2所示：

* 非常明显可以看出销量小于2000的有798例，占到了80%左右，所以大部分香水销量应该集中在2000以下，2000以上的销量就是高销量了。
* 评价100000以上的为菲拉格慕香水。
* 同样在之后作图时候需要用对数处理，这样可以更好查看图的样子。

### 产地

不同产地可以做出产地的饼图来反映市场上销售最多的。首先去除缺失值，做出饼图。

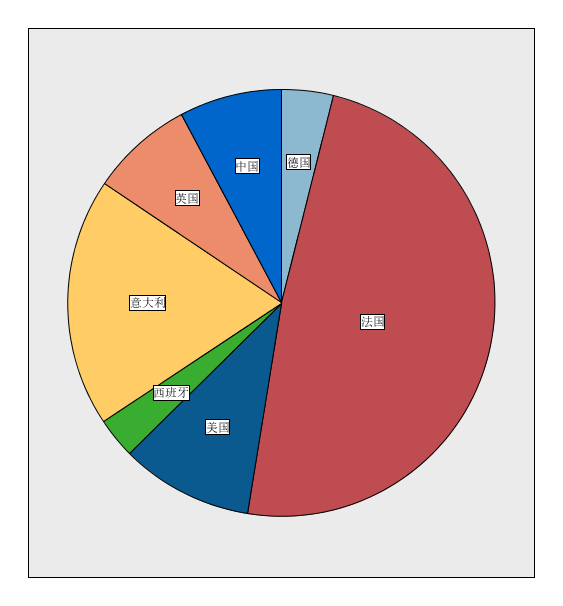


图 2‑3

如图2-3所示：

* 法国是香水生产大国，其次是意大利美国。

# 数据预处理

## 相关性分析

### 商品产地和价格

不同商品产地和价格会有关系也能看出高价区和低价区。

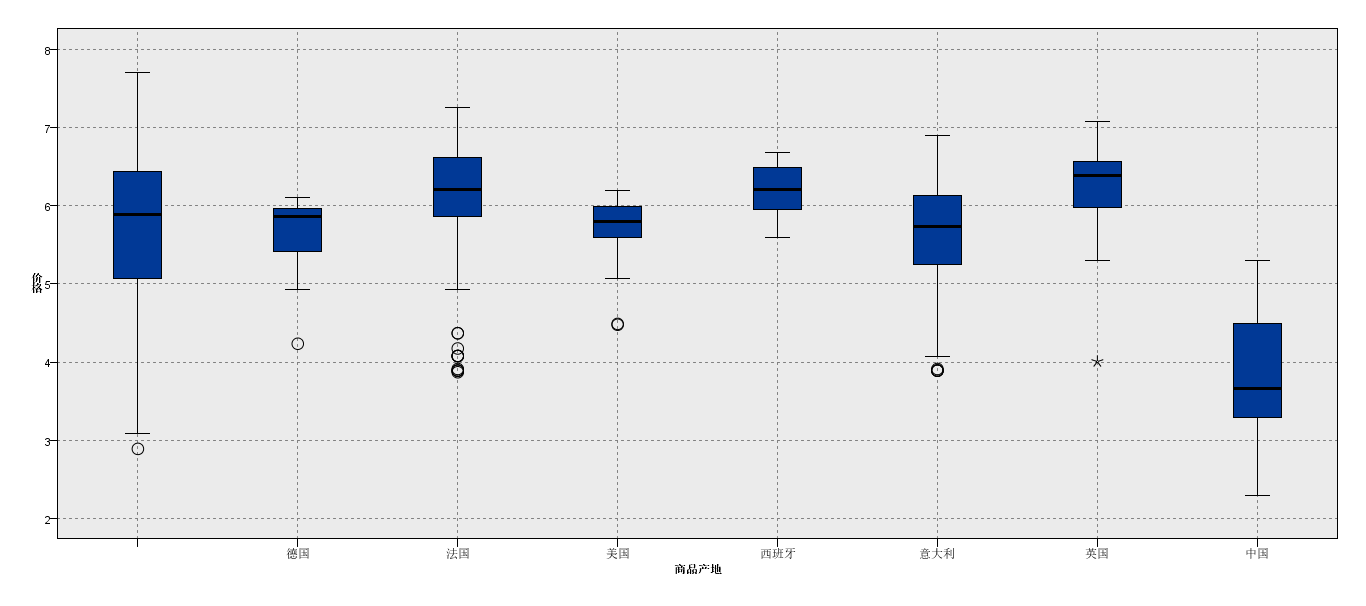


图 3‑1

如图3-1所示：

* 中国香水价格非常低是低价区，法国和英国价格相对较高一些。

### 性别和分类

通用型和女生用的分类很有可能会不同，用分布图作图。

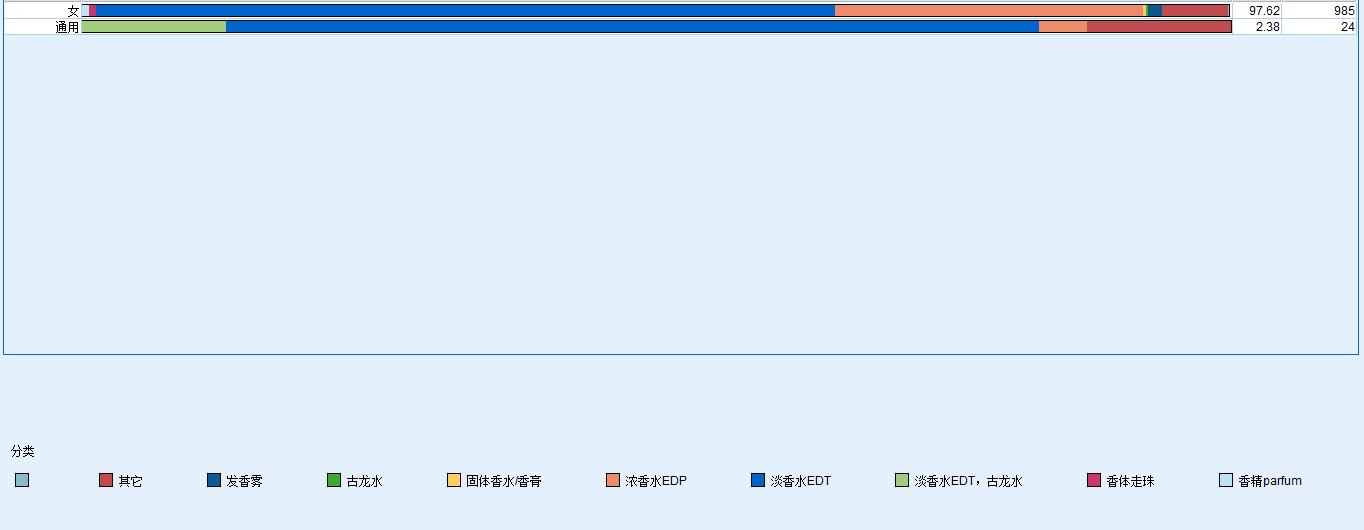


图 3‑2

如图3-2所示：

* 只有通用型的才有古龙水类别，对于女性，没有古龙水分类的

### 毛重和净含量的关系

毛重和净含量本来就是很相近的概念，可能会有冗余，做一下相关性分析。因为这两个变量一开始都是离散的，也不好连续化，所以先采用矩阵进行分析。



图 3‑3

如图3-3所示：

* 毛重和净含量有强相关性，他们独立的概率为0.
* 所以进模型的时候，可以考虑使用净含量来进入，因为净含量的单位更加统一。

### 分类和评价的关系

不同的分类销量肯定不同，可以试着找一下哪种分类销量高，哪种分类销量比较低。

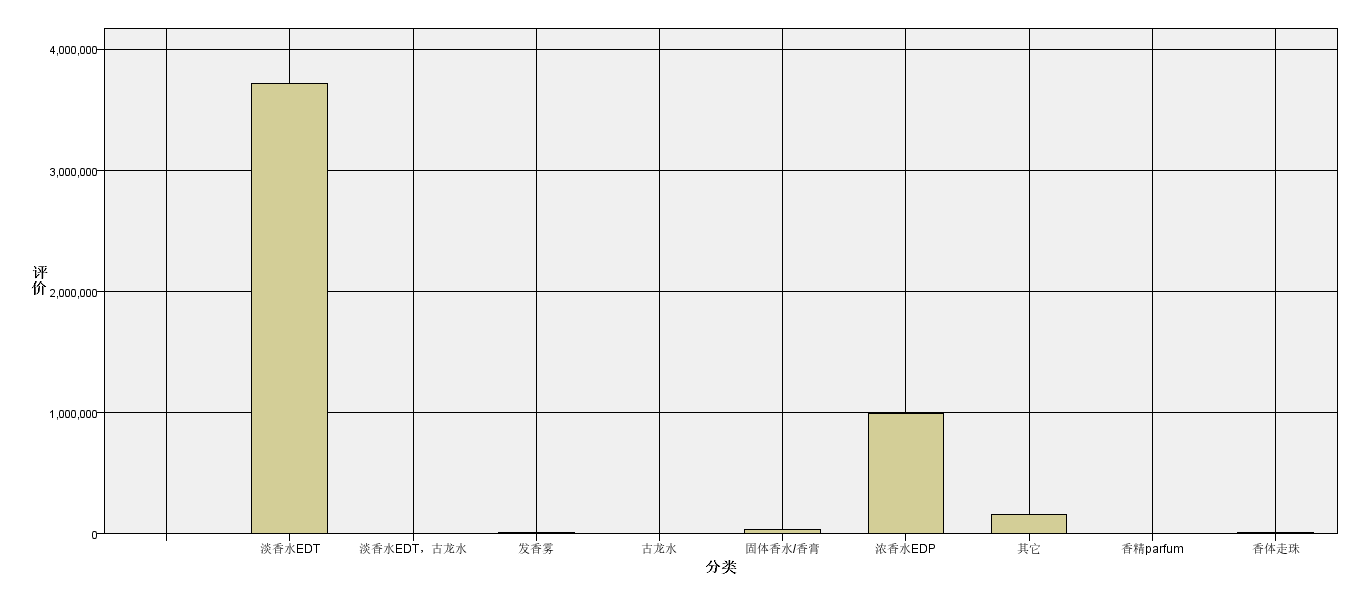


图 3‑4

如图3-4所示：

* 非常明显，市场上主要份额是两种分类香水，一个是淡香水EDT，一个是浓香水EDT。尤其是淡香水EDT销量很可观。
* 如果最后实在没什么可以推荐，可以尝试使用淡香水EDT。

### 商品产地和评价的关系

使用箱图观察产地和评价的关系。

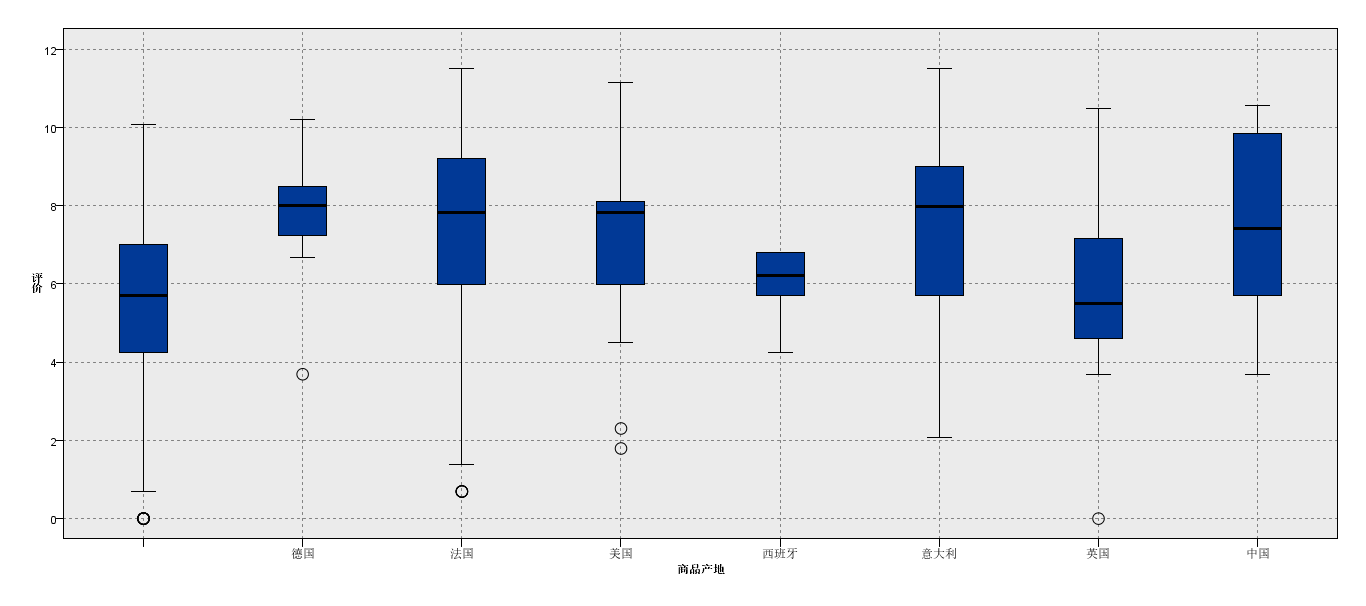


图 3‑5

如图3-5所示：

* 西班牙和英国的稍微低一些，其他的都差不多。

### 包装和评价的关系

香水有时候要送人，或者自己用，不同包装可以满足不同的需求，评价销量或许不同。

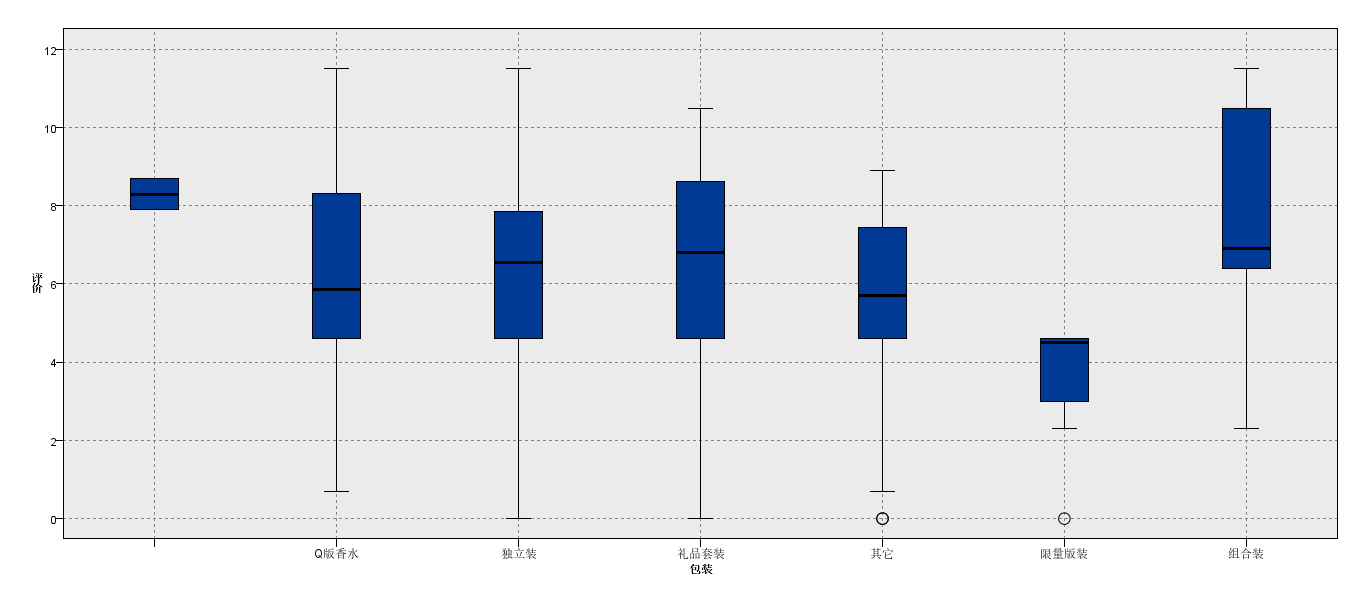


图 3‑6

如图3-6所示：

* 独立装和Q版的评价跨度非常大，可能是因为这两种包装的某种牌子非常受欢迎，造成离群值比较多。
* 组合装的离群值很少，且评价也还算较多，就是中位数以上的销量比较散，可能是受到不同品牌或者价格影响。

### 适用场所的相关性分析

不同的适用场所可能会有相应的关系。因为原来的都是0和1型，要用重新分类结点转化为T/F型才可以。然后输出网络图。

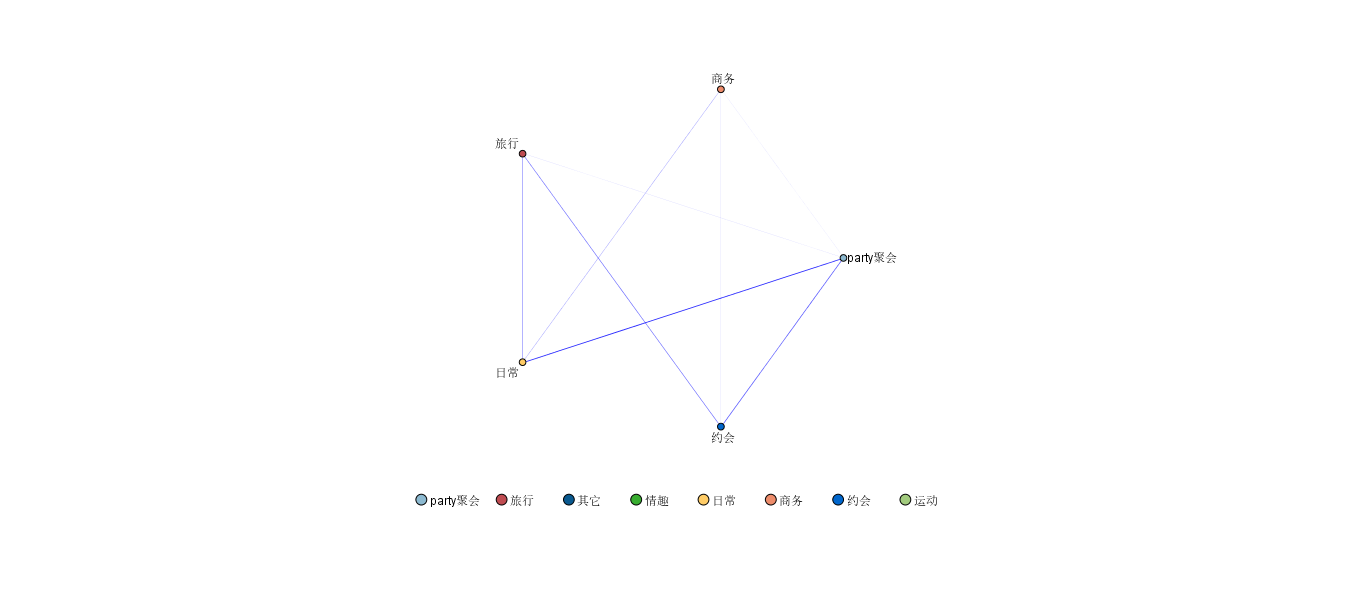


图 3‑7

如图3-7所示：

* 约会和日常经常同时出现。
* 约会、日常、旅行、party聚会和商务经常同时出现。

为了进一步挖掘关联关系，我们采用Apriori进行建模。

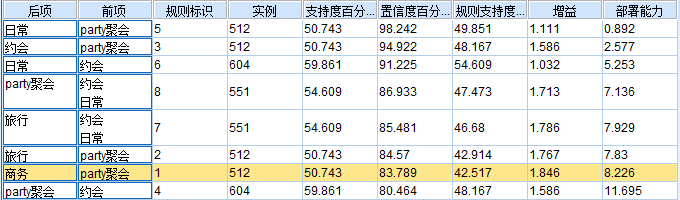


图 3‑8

从图3-8可以看出：

* 日常和Party聚会以及约会经常同时出现，置信度达到87%左右。而且部署能力很强，

### 适用场所、香调和分类之间的关系

不同场所很有可能会用不同的香调和分类的香水，非常值得探索。很遗憾的是，并没有发现特别的东西。

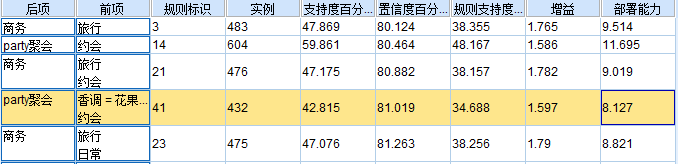


图 3‑9

如图3-9所示：

* 香调为花果香调的用于约会和聚会的比较多，但是由于花果香调的品牌本来就大，并不能说明这个问题。

### 适用场所数量

因为有些香水使用场所数量各有不同，所以考虑使用场所数量会不会与销量有关。导出新字段适用场所数量，因为数量比较小又是整数，所以将适用场所数量直接变成分类属性，然后利用直方图输出查看。

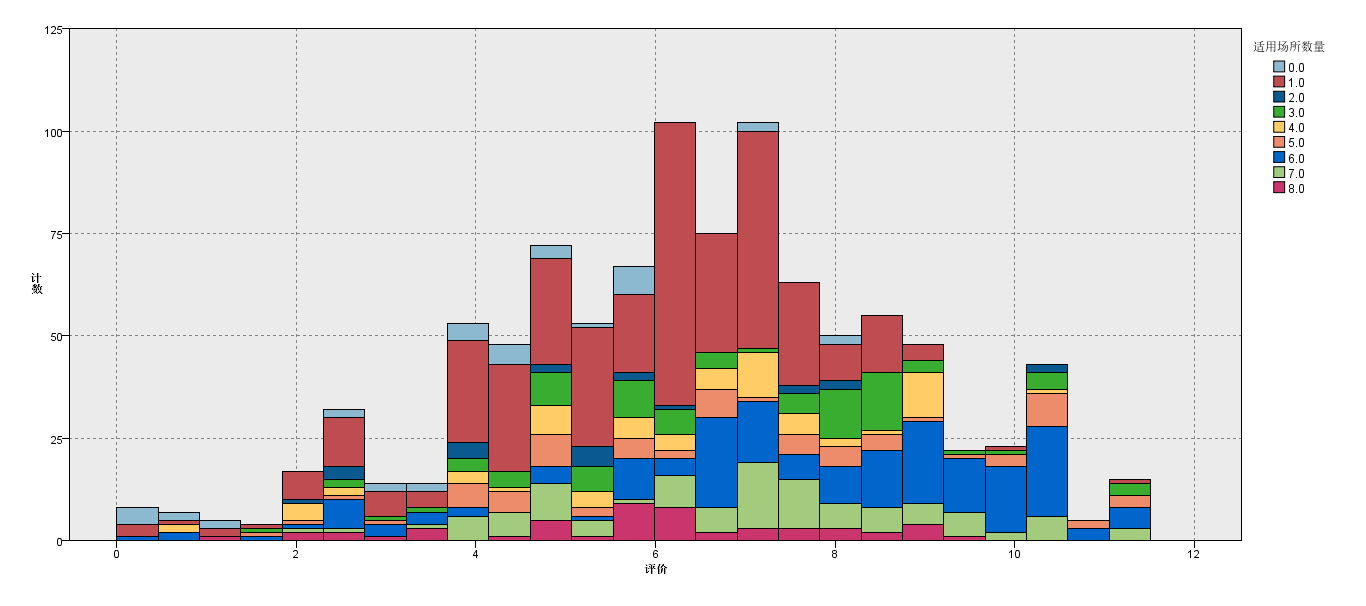


图 3‑10

如图3-10所示：

* 高价格的香水使用范围都还比较广。

## 缺失值填充

对于包装缺失值只有2个，直接用独立包装来填充。香调也使用花果香调来填充。净含量仅有2个缺失值，直接找出来，用相同毛重的东西来填充，香奈儿是31-100ml，古驰填充1-15ml。产地太多，暂时不填充。

## 离散化数值

价格和评价都是连续数值，有必要分成离散的数值进行操作。

### 价格离散化

通过对行业的了解，我们对香水的价格进行离散化。100一下为低，100-300为较低，300-500为中等，500-700较高，700以上为高。



图 3‑11

### 评价的离散化

同样是通过对行业的理解，将评价离散化为销量等级。100以下非常低，100-500为低，500-1000为较低，1000-2000为中等，2000-5000较高，5000-10000高，10000以上非常高。



图 3‑12

离散化后将销量等级和价格等级做相关性分析，利用矩阵进行分析。



图 3‑13

如图3-13所示：

* 销量等级和价格等级有关系，说明离散化还是正确的。

# 建立模型

将销量等级作为目标变量，商品产地，价格等级，性别，净含量，适用场所，适用场所数量，香调，香料作为输入变量，建立C5.0模型。去除商品产地缺失和包装缺失值的影响。

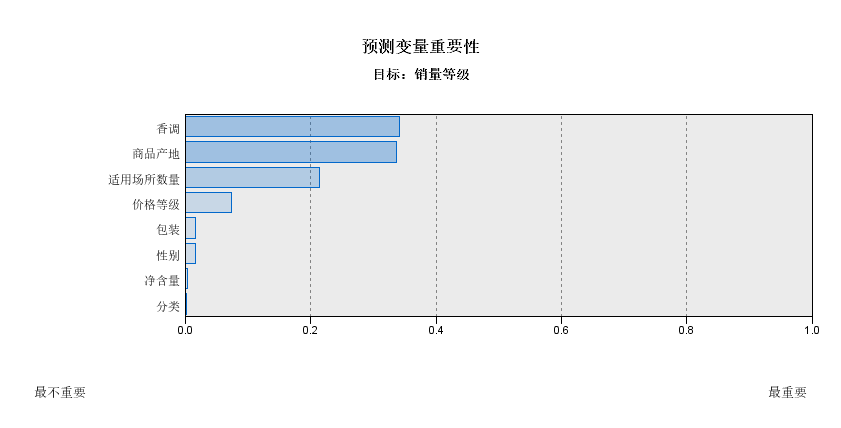


图 4‑1

对销售影响最大的是香调，其次是产地、适用场所数量、价格等级和包装。

# 建议和意见

通过各类分析以及决策树的建立，我们可以得到一些建议和意见：

* 中国和法国的香水销量都不错，可以考虑多进些。
* 英国香水销量比较低，少进一些。
* 买产地在中国的香水时候，多进花果香调，混合香调和花果香调中的分类在浓香水EDP的香水以及东方香调的香水，这样子的销量就会更加高。
* 对于法国的香水，适用场所要比较多的才比较好销售，最好适用4个场所以上。
* 法国香水，木质香调的淡香水EDT销量好，可以多进货。
* 淡香水EDT的销量与价格呈现负相关，而浓香水EDP的销量与价格呈正相关，说明浓香水的买家比较重视品质，对价格不敏感，而淡香水EDT的买家对价格敏感，因此对于不同类型的产品要有具体的价格策略。