**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？**

**我们以从某电商网站上抓取到的香水产品销量数据（文件名为：香水.XLSX）），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。**



**在这个数据集中，产品的评论数在一定程度上代表了产品的销量，将评价数离散化处理反应“销售等级”，但从上表中可知，“评价”字段包含混合的数字、中文及字符，因此需要将其转化为统一的数值型数据，以方便进一步的离散化处理。**

**“适用场所”在数据集中为文本形式，也不利用进一步的分析，应将其分解成“旅行”、“约会”、“其他”等8个字段，根据其是否出现将其值设为“0”或“1”，同时可生成新特征“适用场合数量”。**

**“商品场地”中也存在着地址范围不一致，将“中国大陆”、“广东”等统一替换为“中国”。**

**为了便于大家后续数据分析工作的展开，上述处理我已完成，处理后的文件名为：香水1.xlsx，原文件我也已提供，感兴趣的同学这部分工作也可自行完成。**

**若你面对数据集，毫无分析思路，请继续往下面看，我会给出提示。当然我更希望你先不要看下面的内容，你可以试着按自己的理解去分析，因为它可能会局限你的思路。**

**分析思路：**

**首先确定这是一个描述性任务，找出影响香水销售的影响因素，从而也就可以为香水销售商判定采购计划以及为用户选择香水提供依据。**

**描述性任务的解决方法：相关性分析（统计分析）、关联规则分析、基于规则的分类、聚类分析。**

**(1) 香水销售数据理解、统计分析：了解各个属性值的分布，是否有缺失值、噪声值等，两两属性的相关性分析等。**

**(2) 影响香水销量的因素分析：各个输入特征与目标特征的相关性分析，多个输入特征与目标特征的相关性分析，分析影响香水销量的影响因素，并为特征选择提供依据。**

**(3) 香水适用场所关联分析，试图找出香水适用场所间的关联关系。**

**(4) 基于规则的分类方法即是分类预测同时由于规则的生成又具有很好的描述作用。**

**(5) 香水聚类分析，利用聚类生成簇，分析每个簇的特征。**

**(6) 给出香水营销建议**

**下面的两份目录，仅供参考！**

目录

[第1章 引言 4](#_Toc23429)

[1.1案例背景 4](#_Toc21069)

[1.2任务理解 4](#_Toc22238)

[第2章 数据理解 4](#_Toc17249)

[2.1原始数据表理解 4](#_Toc12388)

[2.2 数据审核与数据分布 5](#_Toc11814)

[第3章 数据预处理与数据准备 7](#_Toc4399)

[3.1相关性分析（两两变量之间） 7](#_Toc4885)

[3.1.1除目标变量之外的两两属性分析 7](#_Toc3212)

[3.1.2其他变量与目标变量的相关性分析 15](#_Toc25117)

[3.1.3适用场所的关联分析 19](#_Toc7485)

[3.2 缺失值填充 19](#_Toc29525)

[3.2.1 商品产地 19](#_Toc27087)

[3.2.2 包装、香调、分类 19](#_Toc11917)

[3.2.3 净含量 20](#_Toc28803)

[3.2.4 适用场所 20](#_Toc8193)

[3.3 类型转换 20](#_Toc32669)

[3.3.1 适用场所转换为二元属性 20](#_Toc27635)

[3.3.2 价格规范化与离散化 21](#_Toc17692)

[3.3.3 评价取对数与离散化 22](#_Toc24591)

[3.4 特征生成与分析 23](#_Toc22813)

[3.4.1 适用场所数 23](#_Toc6318)

[3.4.2 商品名称泛化为品牌 24](#_Toc126)

[3.4.3 有无优惠 25](#_Toc8568)

[3.4.4 产地 25](#_Toc4476)

[3.5 特征选择 25](#_Toc29699)

[第4章 数据分析 25](#_Toc29284)

[4.1 探索性分析（多个变量结合分析） 26](#_Toc3154)

[4.1.1 分类-香调-适用场所数的相关性分析 26](#_Toc10886)

[4.1.1 商品产地-价格-销售等级的相关性分析 26](#_Toc19044)

[4.2 建模分析 27](#_Toc12)

[4.2.1 C5.0模型建模分析 27](#_Toc20793)

[4.2.2 聚类分析 28](#_Toc12659)

[4.3 结论与建议（针对商家或客户） 30](#_Toc22258)

[第5章 问题反馈与心得体会 31](#_Toc4003)

[5.1 问题反馈 31](#_Toc14068)

[5.2 心得体会 32](#_Toc9731)

目录

[1 任务理解及思路整理 1](#_Toc17929)

[1.1 任务理解 1](#_Toc2467)

[1.2 思路整理 1](#_Toc18214)

[2 数据理解 1](#_Toc32475)

[3 数据准备 3](#_Toc11119)

[3.1 特征生成 3](#_Toc32069)

[3.1.1 品牌 3](#_Toc18688)

[3.1.2 适用场所转换为二元属性 4](#_Toc2633)

[3.1.2 适用场所数量 4](#_Toc24684)

[3.2 缺失值处理 4](#_Toc14775)

[3.2.1 包装、香调、净含量、分类 4](#_Toc23979)

[3.2.2 适用场所 5](#_Toc22917)

[3.2.3 商品产地 5](#_Toc2379)

[3.3 离散化 5](#_Toc16555)

[3.3.1 毛重离散化 5](#_Toc29795)

[3.3.2 价格离散化 6](#_Toc485)

[3.3.3 评价离散化 6](#_Toc31956)

[3.4 地区重新分类 6](#_Toc2387)

[3.5 评价数范围缩小 7](#_Toc29919)

[3.6 低频合并 8](#_Toc9301)

[3.6.1 品牌低频合并——小众品牌 8](#_Toc4118)

[3.6.2 产地低频合并 9](#_Toc11544)

[3.6.3 香调低频合并 9](#_Toc2355)

[3.6.4 分类低频合并 9](#_Toc8106)

[4 数据探索 10](#_Toc20323)

[4.1 连续属性与评价的相关性分析 10](#_Toc19599)

[4.2 标称属性与评价的相关性分析 10](#_Toc16117)

[4.3 冗余字段处理 15](#_Toc15667)

[4.3.1 适用场所数量与各场所之间的相关性分析 15](#_Toc18587)

[4.3.2 毛重、净含量相关性分析 15](#_Toc16553)

[4.4 部分标称属性两两与评价的相关性分析 16](#_Toc17310)

[4.4.1 品牌、商品产地与评价e对数的相关性分析 16](#_Toc2961)

[4.4.2 品牌、净含量与评价e对数的相关性分析 16](#_Toc19118)

[4.4.3 分类、香调与评价e对数的相关性分析 17](#_Toc27690)

[4.4.4 包装、净含量与评价e对数的相关性分析 17](#_Toc16436)

[4.4.5 适用场所与评价e对数的关系 18](#_Toc23497)

[5 适用场所间的关联分析 20](#_Toc15964)

[6 场所-品牌关联分析 21](#_Toc16732)

[7 模型构建——影响香水销量的因素分析 21](#_Toc6182)

[7.1 特征选择 21](#_Toc5272)

[7.2 模型构建 22](#_Toc31807)

[7.2.1 未过滤“其他”或“未知”记录的模型分析 22](#_Toc20484)

[7.2.2 过滤“其他”或“未知”记录的模型分析 23](#_Toc3960)

[7.2.3 “小众品牌”记录的模型分析 24](#_Toc20487)

[7.3 模型综合分析 25](#_Toc17420)

[8 香水聚类分析 26](#_Toc28565)

[9 香水营销建议 28](#_Toc7199)

[10 问题与收获 29](#_Toc30526)