一、实验要求

对wine reviews数据集进行频繁模式和关联规则挖掘，数据分析要求如下：

1. 对数据集进行处理，转换成适合进行关联规则挖掘的形式； 
2. 找出频繁模式； 
3. 导出关联规则，计算其支持度和置信度; 
4. 对规则进行评价，可使用 Lift、卡方和其它教材中提及的指标, 至少 2 种； 
5. 对挖掘结果进行分析；
6. 可视化展示

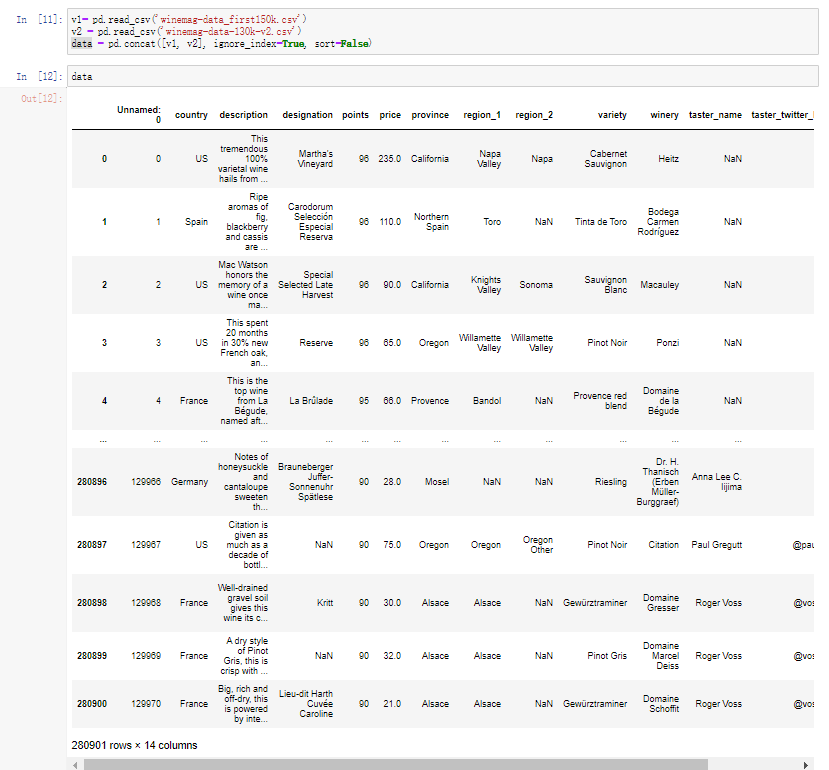
二、wine reviews数据集

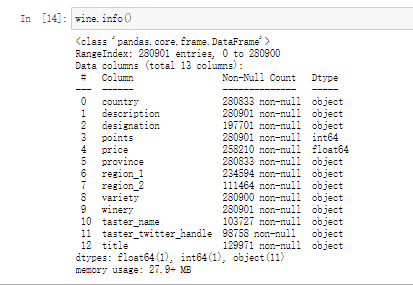
该数据集包含两个数据子集，分别是2017年6月15日爬取的150,000条数据，和2017年11月22日爬取的拥有更丰富信息的130,000条数据。每条数据包含{ 'country', 'description', 'designation', 'points', 'price', 'province', 'region\_1', 'region\_2', 'variety', 'winery'}共十个属性。

属性{'country', 'designation', 'province', 'region\_1', 'region\_2', 'variety', 'winery'}，分别表示国家名、葡萄园名、省名、产区1名、产区2名、葡萄类型、酿酒厂名，这些都是标称属性。

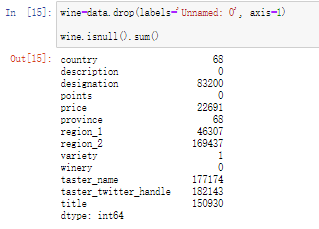
属性{ 'points', 'price'}，分别表示每条数据对葡萄酒的评分以及这瓶葡萄酒的成本，这些是数值属性。

为了更好地进行挖掘，我们首先查看了数据集的大致信息如下所示：





接下来，我们对数据集进行了预处理，删掉数据集中的缺失值，缺失值信息如下图所示：



最后我们需要选择合适的属性进行频繁模式和关联规则挖掘，选择的属性如下: country、points、price、 provin、region\_1、region\_2、variety 以及 winery，将预处理后的数据集中的这些属性字段提出出来，并放在一个列表里，作为后续频繁模式挖掘和关联规则挖掘的数据对象。

三、频繁模式

频繁模式：是频繁得出现在数据集中的模式；频繁项集导致发现大型事物或者数据集之间有趣的关联或相关性，频繁项集挖掘的一个典型案例是购物篮分析,该过程通过发现顾客放入他们的购物篮中的商品之间的关系，分析客户的购物习惯。为了更好的进行频繁模式挖掘，本次实验选择了注名的Apriori算法，来提升挖掘的效率。

Apriori算法：限制候选产生发现频繁项集，具有先验性质，频繁项集的所有非空子集也一定是频繁的。具体思想如下：

如果项集I不满足最小支持度阈值，即，那么（超集）不可能比项集I更频繁。这种性质又称为反单调性：一个集合不能通过测试，那么它的所有超集都不能通过。

算法流程如下：

给出一个事务数据，Txx是一个事务的数据，lx是对应的项：

T100：{l1,l2,l5}；T200：{l2,l4}；T300：{l2,l3}

T400：{l1,l2,l4}；T500：{l1,l3}；T600：{l2,l3}

T700：{l1,l3}；T800：{l1,l2,l3,l5}；T900：{l1,l2,l3}

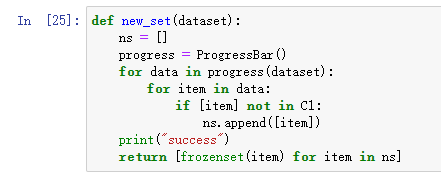
1、项集L1：扫描全部事务，计算支持度计数，判断是否满足最小支持度，不满足的剔除。

2、项集L2：1项集进行全连接，产生2项集集合，具有 [公式] 个，计算2项集的支持度计数，并进行剪枝（先验性质和反单调性特性）。

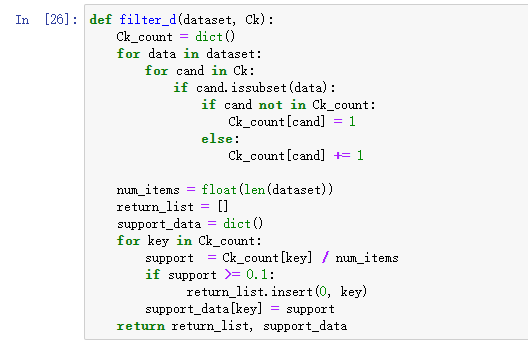
3、项集L3：对2项集进行全连接，并进行剪枝。

4、以此类推，最终找到所有频繁项集。

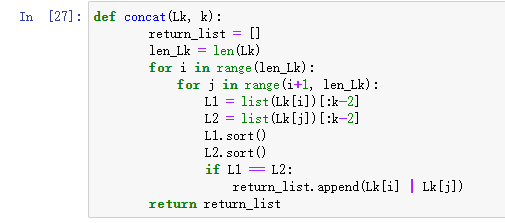
实现 Apriori 算法需要要构建全部可能的单元素候选项集合，以列表的形式存储这些单元素候选项集合，每个单元素候选项包括属性名，属性取值，具体过程如下所示：



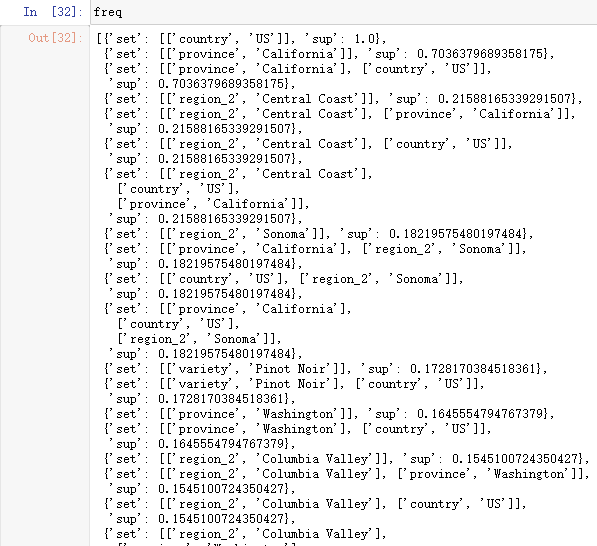
然后，需要过滤掉小于最小支持度的非频繁项集，判断每个候选项集中元素的出现频率， 过滤掉出现频率低于最小支持度的项集，最小支持度设置为 0.1。 如下所示：



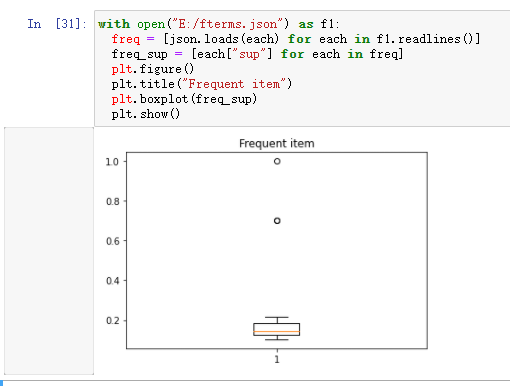
最后，对所有的候选集进行合并及筛选：



通过Apriori算法计算后得到挖掘结果，按支持度从大到小排序，具体的结果如下图所示。从图中我们可以看到 country 和 US 同时出现， 原因可能是在数据预处理中删除了大量包含空值的元组，导致剩下的数据中 country 属性只剩下了美国，同时加利福尼亚在省份中出现的频率最高，表示数据中大部分就出自加利福尼亚州：



以盒图的形式展示结果如下：



四、关联规则

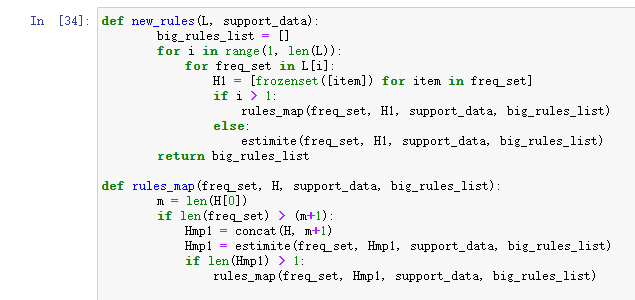
关联规则是发现交易数据库中不同商品（项）之间的联系。这些规则找出顾客购买行为模式，如购买了某一商品对购买其他商品的影响。具体的挖掘过程如下：

1、找出所有的频繁项集，该项集频繁出现的次数至少在最小支持度阈值计数以上。

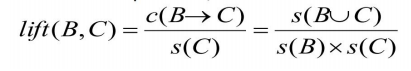
2、由频繁项集产生强关联规则，项集同时满足最小支持度和最小置信度阈值

对于使用 Apriori 算法进行频繁模式挖掘所得到的频繁 项集计算其关联规则并用评价指标进行评价。本实验利用 Lift 和 Jaccard 两种 指标进行评价。

基于 Apriori 算法，首先从一个频繁项集开始，创建一个规则列表，规则右 部只包含一个元素，然后对这些规则进行测试，接下来合并所有的剩余规则列表 来形成一个新的规则列表，其中规则右部包含两个元素；同时，我们还需要将频繁项集的元素映射到规则右面的元素列表：



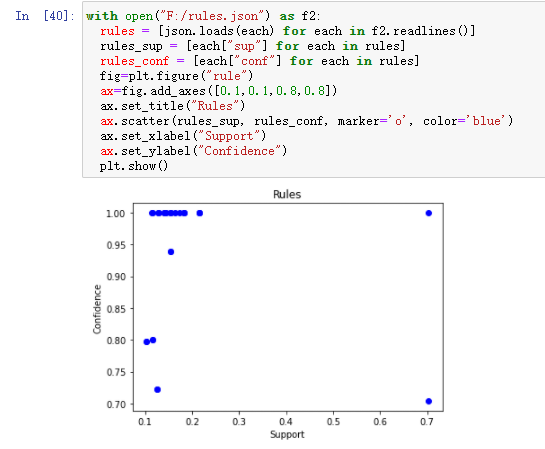
Lift 评价指标的计算公式如下：



Jaccard 评价指标：



利用相关的评价指标来评估生成的规则，结果绘制的散点图如下所示：



然后，我们需要将得到的关联规则以及评价结果按置信度排序，由于 country 字段的取值 全部为 US，所以对齐分析没有实际意义，因此跳过 country 字段，通过结果可以

看出 Region\_2 和 province 的关联度较高，表示该地区很可能在对应的省内，如下图所示

