智能数据挖掘作业8

19200300004 黄铭瑞

# 实验目的

使用apriori算法对数据进行关联分析。

# 实验原理

## Apriori算法

Apriori算法是常用的用于挖掘出数据关联规则的算法，它用来找出数据值中频繁出现的数据集合，找出这些集合的模式有助于我们做一些决策。

### 关联分析

找出频繁一起出现的物品集的集合，我们称之为频繁项集，在频繁项集的基础上，使用关联规则算法找出其中物品的关联结果。

### 支持度

支持度就是几个关联的数据在数据集中出现的次数占总数据集的比重。或者说几个数据关联出现的概率。

### 置信度

置信度体现了一个数据出现后，另一个数据出现的概率，或者说数据的条件概率。

### 提升度

提升度表示含有A的条件下，同时含有B的概率，与X总体发生的概率之比。

## Apriori的两个定律

定律1：如果某商品组合小于最小支持度，则就将它舍去，它的超集必然不是频繁项集。

定律2：如果一个集合是频繁项集，即这个商品组合支持度大于最小支持度，则它的所有子集都是频繁项集。

## 算法步骤

1. 扫描整个数据集，得到所有出现过的数据，作为候选频繁1项集。k=1，频繁0项集为空集。
2. 挖掘频繁k项集
3. 扫描数据计算候选频繁k项集的支持度。
4. 去除候选频繁k项集中支持度低于阈值的数据集,得到频繁k项集。如果得到的频繁k项集为空，则直接返回频繁k-1项集的集合作为算法结果，算法结束。如果得到的频繁k项集只有一项，则直接返回频繁k项集的集合作为算法结果，算法结束。
5. 基于频繁k项集，连接生成候选频繁k+1项集。
6. 令k=k+1，转入步骤(2)。

# 实验过程

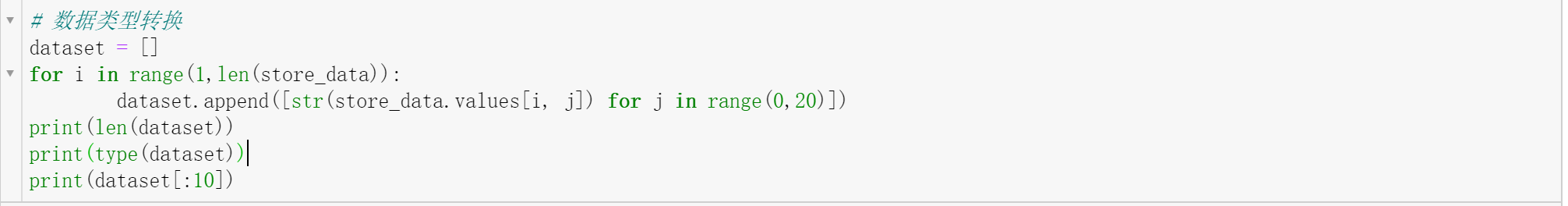
## 读取数据

Pd.read\_csv()读取商品关联数据。



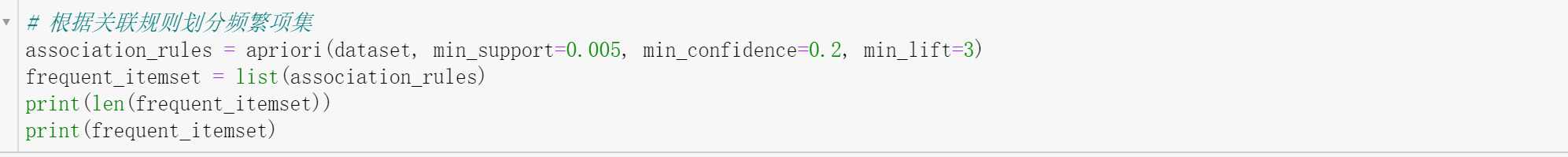
## 类型转换

把数据转为列表

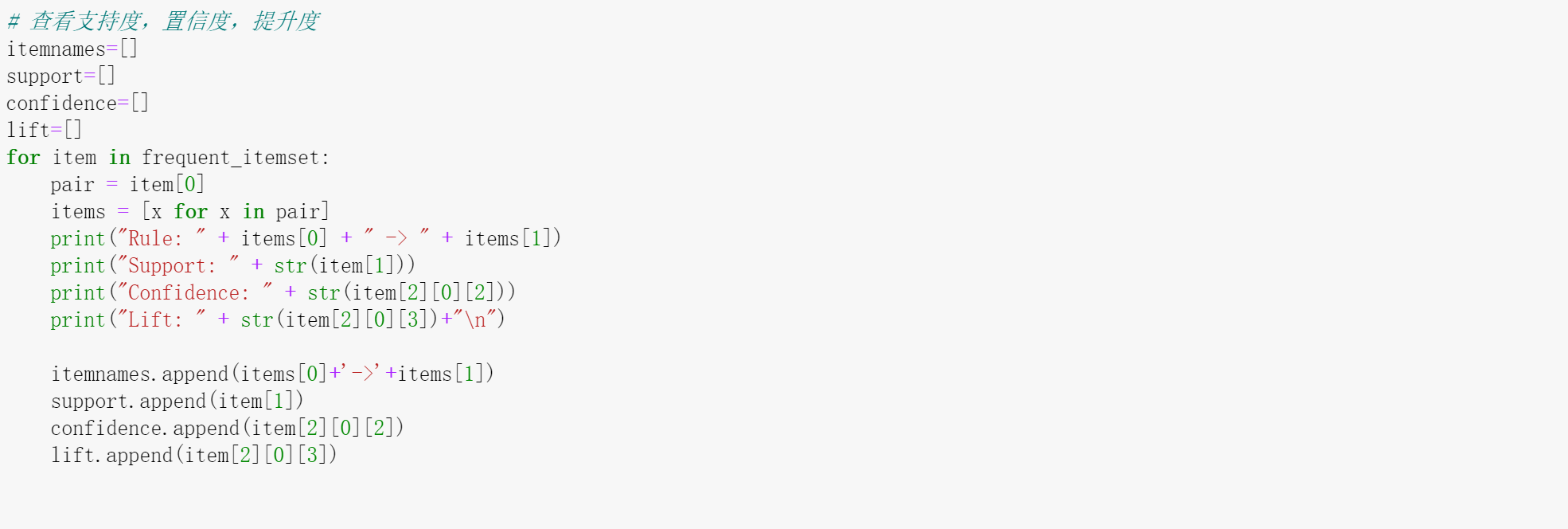


## 频繁项集

from apyori import apriori，使用apriori（）给定好最小支持度为0.005，最小置信度为0.2，最小提升度为3。



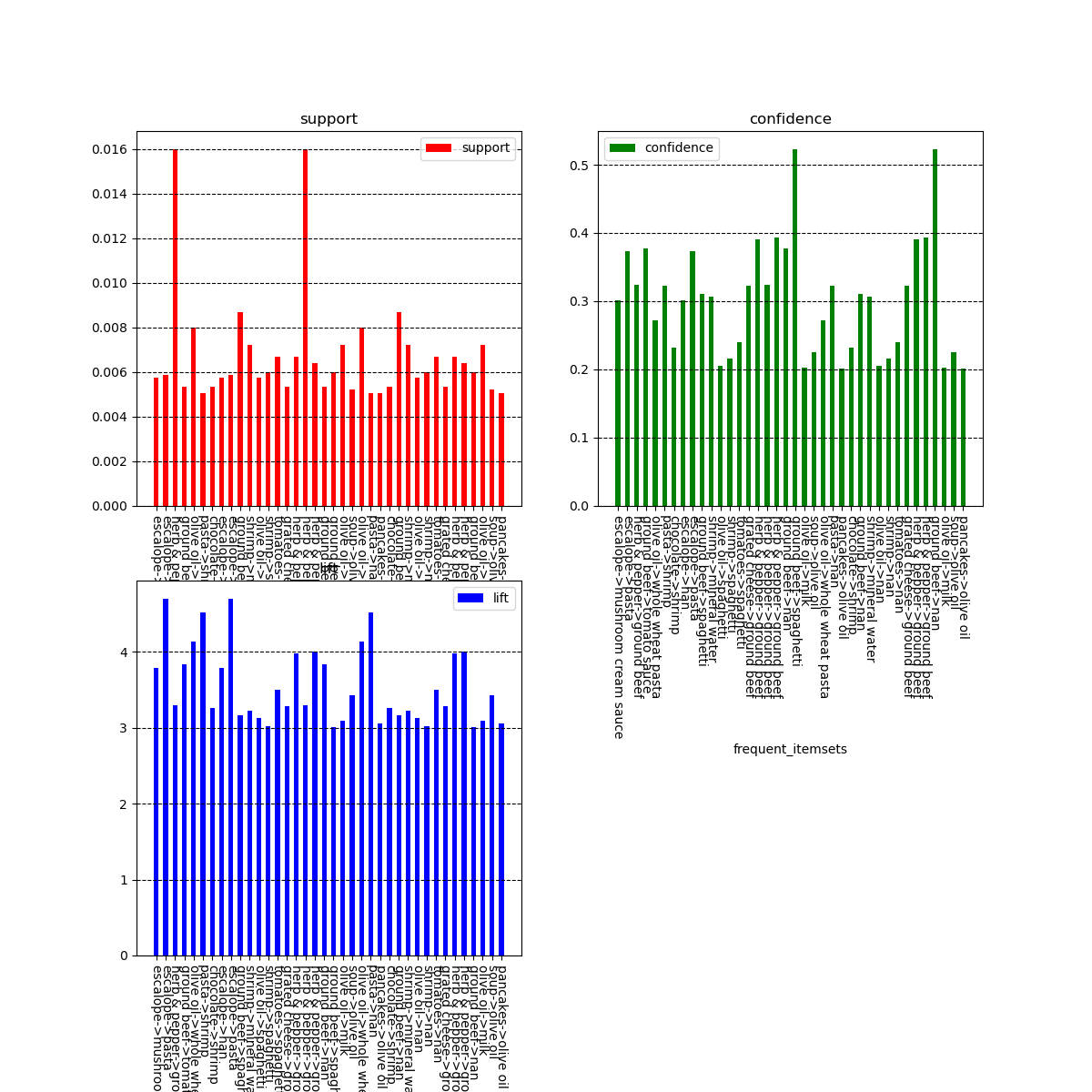
查看各频繁项集的支持度，置信度，提升度。



作图



# 实验结果



# 附录

<Apriori.html>