2022 学年第一学期
《数据挖掘原理与技术》
期末大作业

报告题目：基于关联规则挖掘的电源适配器市场交易数据分析

姓名：严珏

学号：20307100120

专业：管理学系

年级：大三

时间：2023 年 1 月

目录

1. 研究背景

1.1 研究背景

随着互联网的发展和电子商务的普及，网购已经成为了人们主要的购物渠道，各个电子商务网站也蓬勃发展。电子商务拉近了商家与消费者的距离，使得商家与消费者的沟通更加直接而高效，通过对电子商务网站上收集的数据进行分析，也让卖家更加方便地洞察消费者的偏好、以及业内其他竞争者的动向。

如今，电子产品已经充斥了我们的生活，在日常的生活中，手机，平板，笔记本电脑的使用量正在逐步的增加，同样的，电源适配器这类 3C 配件的对市场也在不断扩大。

通过对某电子商务网站上一批“电源适配器 (Power Adapter)”卖家的交易数据进行分析，既能够帮助卖家了解电源适配器市场中的消费者行为偏好，也能够比较不同商家间的表现和差异。

1.2 研究设计

基于上述背景，本课题拟研究如下问题：

从该电子商务网站所有商家构成的电源适配器市场（后文简称“电源适配器市场”）出发：

1) 市场中，哪些类型的产品最畅销？

2) 消费者喜欢同时购买哪些类型的产品？

从电源适配器卖家的角度出发：

1) 哪些卖家表现最好，他们的特点是什么？

2) 卖家可以怎么做来提高自身的销量？

2. 研究框架

2.1 研究思路

1) 消费者行为研究：

从整个电源适配器市场出发，对数据集进行关联规则挖掘，研究市场的消费者行为

- 将数据集中产品相关信息整理为产品信息表，并在此基础上对产品进行抽象化处理
- 对抽象化后的“产品类”进行关联规则挖掘

2) 竞争格局研究：

对于单个卖家进行关联规则挖掘，研究市场中商家的竞争格局，及明星商家的成功原因

- 将数据集中商家相关信息整理为商家信息表，找出表现最好的商家
- 对于该商家进行关联规则挖掘

2.2 研究内容

2.2.1 数据清洗

a) 数据集描述：

Transaction 数据集共有 5078 行，以及 5 个变量，该数据集包含字段：卖家 ID（seller_id）、买家 ID（buyer_id）、产品 ID（product_id）、产品售价(RMB)（product_price）、此产品历史总销售量（sold_quantity）

	seller_id	buyer_id	product_id	product_price	sold_quantity
0	1	71	4776	21.0	20.0
1	1	121	2837	23.0	83.0
2	1	554	3248	21.0	44.0
3	1	573	2852	27.0	29.0
4	1	573	4114	25.0	105.0

b) 异常值处理：

- 缺失值：数据集在‘sold_quantity’项有两个缺失值，删除缺失值所在行后，数据集剩下 5076 行
- 异常值：发现对于同一卖家销售的同一商品，会出现价格不同的问题：

	seller_id	buyer_id	product_id	product_price	sold_quantity
4	1	573	4114	25.0	105.0
44	1	1172	4114	21.0	105.0
71	1	1974	4114	25.0	105.0

解决方法：同一商品的价格不同，可能是由于购买的时间不同，商品价格存在起伏，该现象是合理的。在后续处理中，若涉及到价格因素，则以一组（product_id，product_price）标识一件商品。

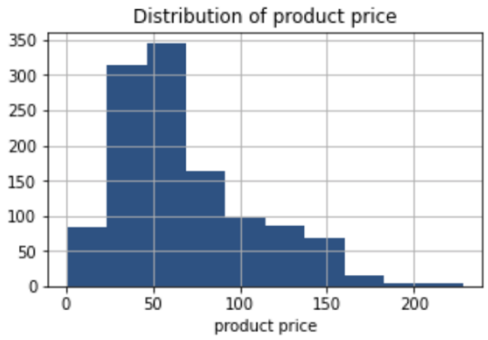
2.2.2 数据预处理

a) 产品信息整理：

将数据集中产品相关信息整理为产品信息表。由于涉及到价格，以一组（product_id，product_price）标识一件商品，即拥有不同价格同一个产品会在产品信息表中存为两列，但由于其本质上是用一件商品，所以销量相同。

将产品按照不同价格和不同销量进行抽象化。

- 价格抽象化：



电源适配器市场中，商品的价格分布在 0-250 元，其中 50 元左右的商品较多。根据其分布，选择价格的 0.9 分位数（120）和 0.4 分位数（49）作为分类的标准：

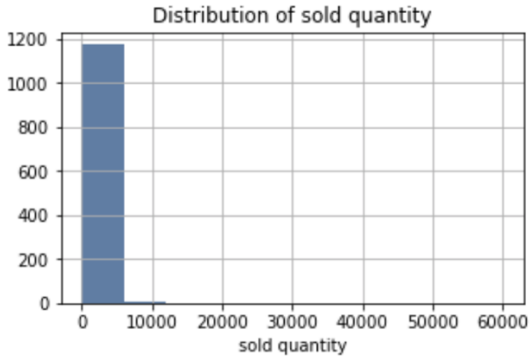
价格>120 元：高端产品（high-end）

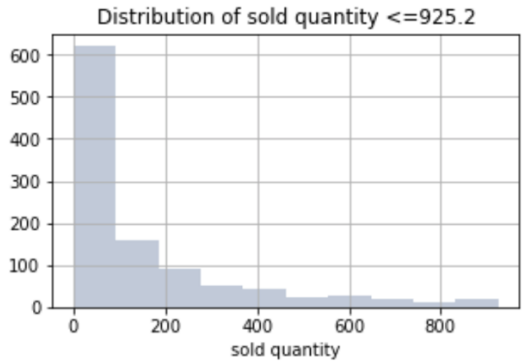
价格在 50 和 120 元之间：中端产品（middle-end）

价格<50 元：低端产品（low-end）

- 销量抽象化：

可以看到，产品销量的分布很不平均，销量的 0.9 分位数为 925.2，这意味着 90%以上的产品集中在 1000 以下，但是个别几个明星商品销量超过 10000，销量最高的商品达到 59967。去掉 outliner 后再次作图查看其分布，可以看到仍是集中在 0-100，销量高的产品较少。





根据上述分布，选择 0.5 分位数（83），0.75 分位数（303.5），0.9 分位数（925.2）作为分类的标准。同时，发现还有销量为 0 的产品，可能是刚刚上架的新产品，将其单分为一类，最终分为 5 类：

销量>925.2：畅销品（best-seller）

销量在 303.5 和 925.2 之间：优质品（well-seller）

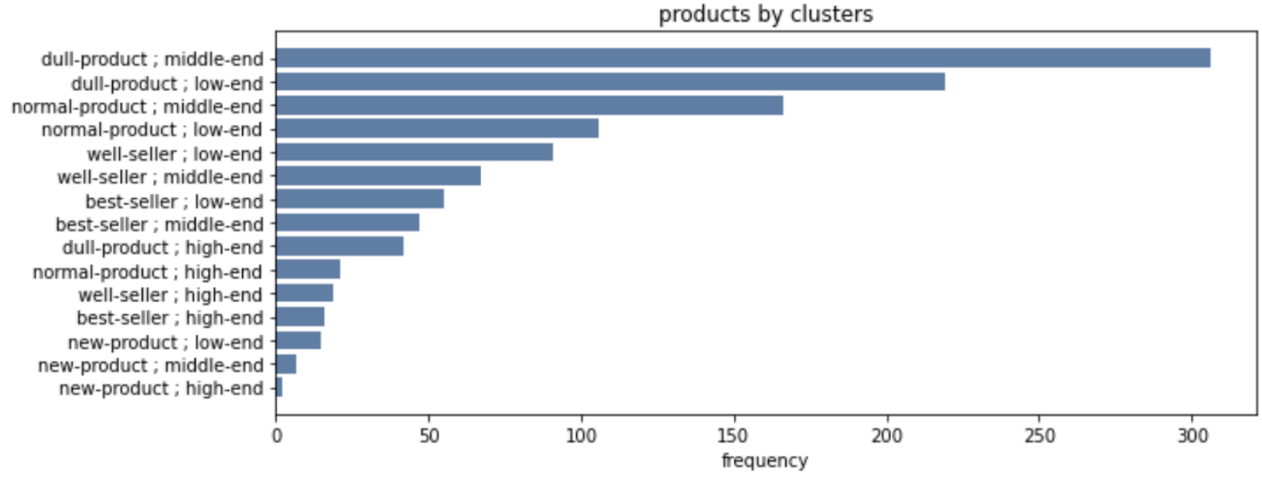
销量在 83 和 303.5 之间：普通商品（normal-product）

销量<83：滞销品（dull-product）

销量=0：新产品（new-product）

- 编码结果：

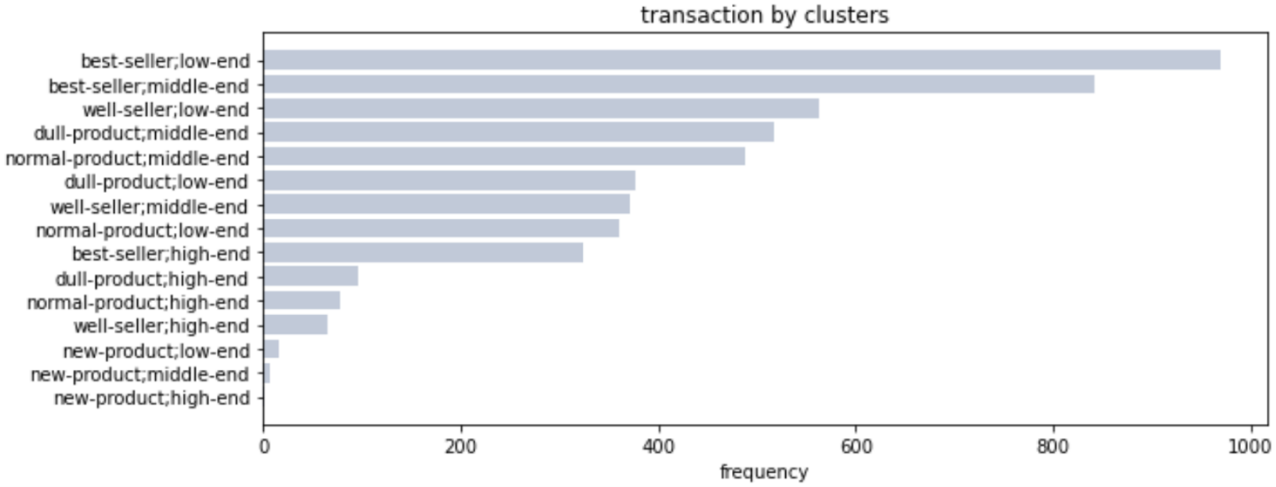
最终将价格和销量组合起来，得到 3*5=15 个产品类，每个类别的商品数量如下：



可以看到，“滞销的中端商品”这一类别的商品数量是最多的，其次是“滞销的低端商品”。从商品的角度看，电源适配器市场中的商品集中在中低端，大部分商品的销售表现一般。

b) 订单信息整理：

- 将上述编码结果应用到 Transaction 数据集中：



对比可以发现，“畅销品”和“优质品”虽然商品类别很少，但是被购买的次数很多。在本交易数据集中，最畅销的品类是“低端的畅销品”和“中端的畅销品”，其次是“低端的优质品”、“中端的滞销品”和“中端的普通商品”。可以认为，大部分消费者会选择购买中低端（即价格在 120 元以下）的电源适配器产品。

- 将 transaction 整理为 order：

以一组(seller_id, buyer_id)识别为一个订单，将这个订单的 id 设定为“‘seller_id’-‘buyer_id’”的形式，再将其整理为一个新的订单表，该订单表包含信息：订单编号（order_id）、该订单包含

的商品 ID（product）、该订单包含的商品类（product_type）

order_id	product_type	product
1-1045	[dull-product;low-end]	[3360]
1-1047	[dull-product;low-end, well-seller;low-end]	[2581, 2587]
1-1054	[dull-product;low-end, normal-product;low-end]	[4112, 4353]
1-1079	[normal-product;low-end]	[2915]
1-1084	[well-seller;low-end]	[2587]

总共得到 3672 个订单，其中包含 1174 个商品（商品 ID）和 15 个产品类。

2.2.3 消费者行为探究

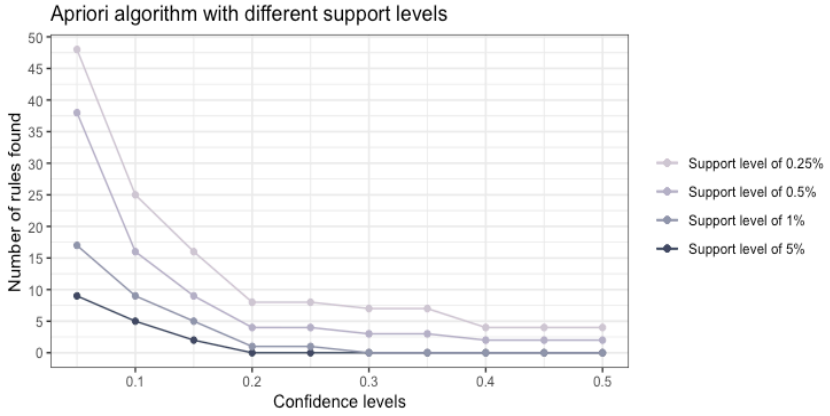
使用抽象后的“产品类”，对于由所有卖家构成的整个电源适配器市场进行关联规则挖掘，探究该市场中的消费者行为偏好。

- 算法选择：

选择 R 语言中的 apriori 算法进行关联规则挖掘。但是由于数据量不大，apriori 算法和 fpgrowth 算法均用时很短，没有显著区别。

- 参数（置信度、支持度选择）：

右图为不同支持度和置信度下，寻找到的关联规则数量，为了避免关联规则过多或过少，选择支持度为 0.5%、置信度为 0.15 作为筛选的标准，筛选得到 9 条关联规则，其中 7 条长度>1，后续将重点分析这 7 条关联规则。

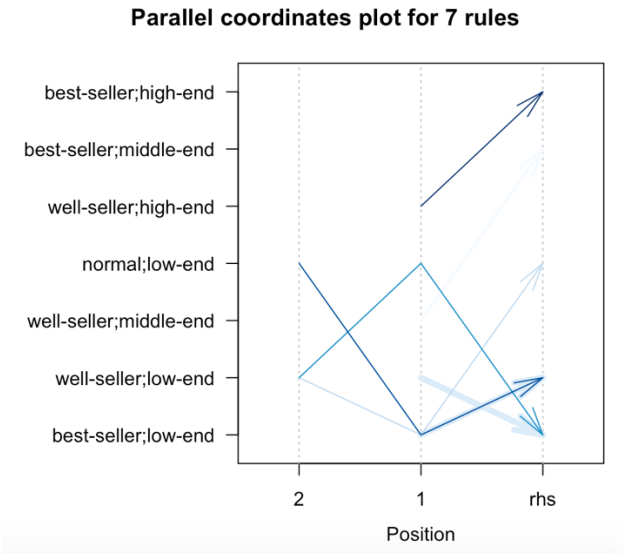


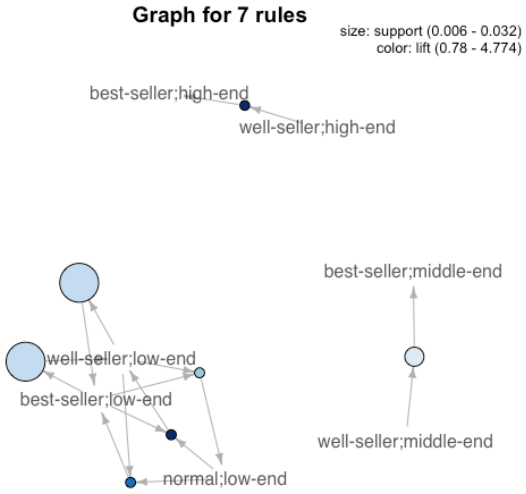
七条关联规则及其指标如下（按照 lift 从大到小排序）：

```
lhs                                rhs                                support confidence    coverage      lift
{well-seller;high-end}          => {best-seller;high-end}          0.005991285  0.3548387 0.016884532 4.7727756
{best-seller;low-end,
 normal-product;low-end}        => {well-seller;low-end}           0.006263617  0.6052632 0.010348584 4.7187395
{normal-product;low-end,
 well-seller;low-end}           => {best-seller;low-end}           0.006263617  0.6388889 0.009803922 3.4500000
{best-seller;low-end,
 well-seller;low-end}           => {normal-product;low-end}        0.006263617  0.1949153 0.032135076 2.1429006
{well-seller;low-end}           => {best-seller;low-end}           0.032135076  0.2505308 0.128267974 1.3528662
{best-seller;low-end}           => {well-seller;low-end}           0.032135076  0.1735294 0.185185185 1.3528662
{well-seller;middle-end}        => {best-seller;middle-end}        0.014433551  0.1531792 0.094226580 0.7801304
```

其中支持度最高的规则来自【“低端的畅销品”和“低端的优质品”】，这说明该电子商务网站的客户大多数会同时购买这两种商品。而 lift 最高的规则来自【“高端的优质品”和“高端的畅销品”】，这说明购买“高端优质品”的消费者很大概率也会购买“高端的畅销品”。

- 关联规则可视化及分析：





从上左图中可以清晰地发现，不同大部分关联规则集中在“低端”系列产品中，且低价格产品相关的规则支持度都较高。这说明来该电子商务网站购买电子适配器的顾客大多会选择低端产品，且购买低端产品的顾客也倾向于购买其他类别的低价格产品。

从上右图中可以发现，不同价格的产品类之间，没有交叉的关联规则。这意味着不同价格带可能针对着不同的消费者群体，购买低价格产品的用户也会同时购买低价格产品，哪怕这些商品的销量没有那么高。而购买高端产品的用户，也会同时购买高端产品。

2.2.4 市场竞争格局探究

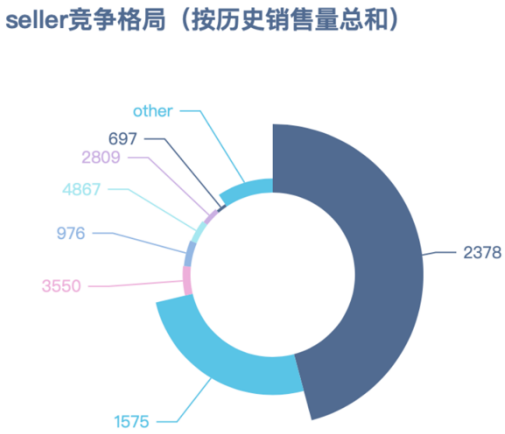
在从整个电源适配器市场的角度，对于产品类进行关联规则挖掘后，我们考虑从商家的角度，对于具体的商品进行挖掘。

- 商家信息整理

将信息按照 seller 进行分类，整理出商家信息表，表格包括信息：商家 ID（）、该商家商品种类数（product_number）、该商家在原数据集中售出的商品总数（sold_transaction）、 该商家所有商品的价格均值（product_price_mean）、该商家所有商品的历史总销量（sold_quantity_sum）

seller_id	product_number	sold_transaction	product_price_mean	sold_quantity_sum
976	83	588	24.323129	788424.0
1575	44	546	60.613553	4267546.0
2378	53	422	137.507109	7686331.0
151	40	376	77.194149	270198.0
2809	70	355	37.539718	540395.0

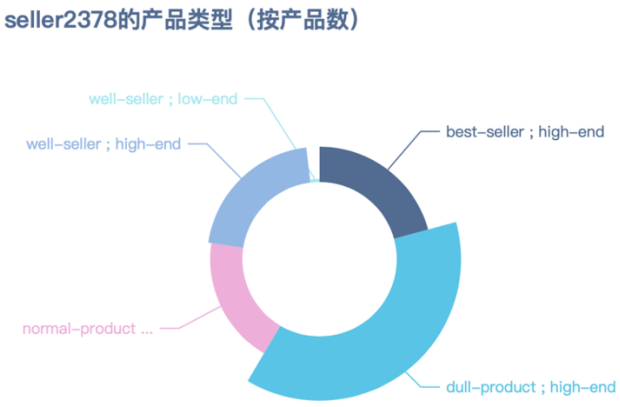
以历史总销量为指标，绘制饼图，查看电源适配器市场的竞争格局（如右图）。

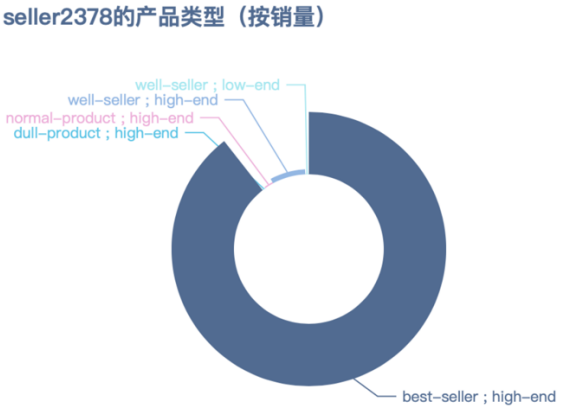


根据右图，发现电源适配器市场的竞争格局集中，头部的两个商家（ID 分别为‘2378’和‘1575’）已经占据了市场份额的 71.28%。接下来，将以占据历史总销售量 45.8%的商家 2378 为例，研究其特点，探究其畅销的原因。

2.2.5 明星店铺案例研究（商家 2378）

1) 商家 2378 的商品特点:





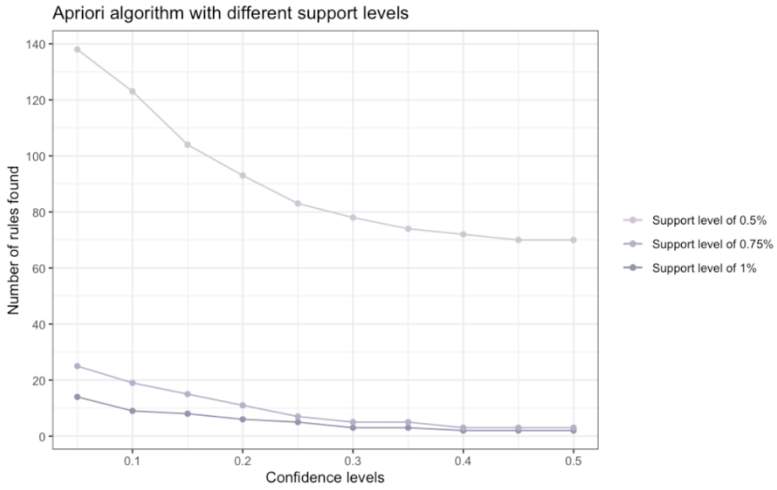
分别以产品数量和产品总销量为指标绘制饼图，可以发现，商家 2378 的商品价格类型集中，53 件产品中有 52 件都属于高端类型，但是有很大一部分销量并不好，有 20 件滞销品。但是商家 2378 的 11 个畅销品贡献了店铺 89.42%的销售量。

2) 商家 2378 的商品关联规则挖掘:

对商家 2378 的所有 310 个订单和 53 件商品进行关联规则挖掘。

- 算法选择：同上，选择 R 语言中的 apriori 算法进行关联规则挖掘。
- 参数（置信度、支持度选择）：

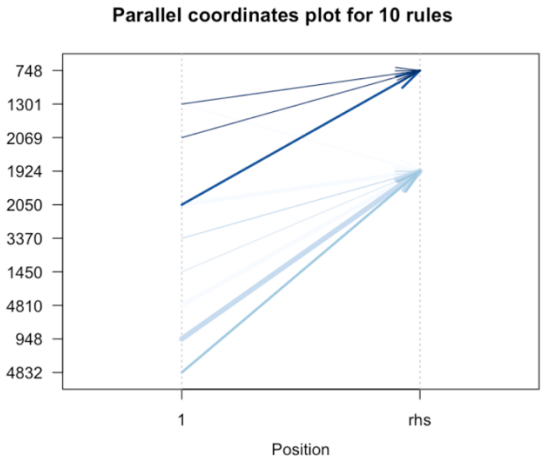
设置不同的置信度和支持度参数，比较算法找到的关联规则数量。可以发现支持度为 0.5%时，找到的规则太多，难以分析。选择支持度为 0.5%、置信度为 0.2 作为筛选的标准，筛选得到 11 条关联规则，其中 10 条长度>1，后续将重点分析这 10 条关联规则。

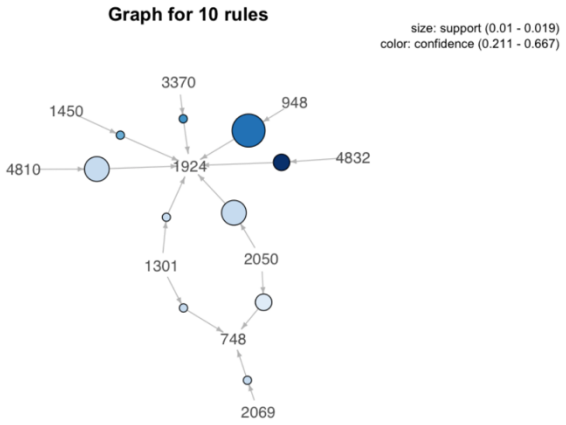


十条关联规则及其指标如下（按照 lift 从大到小排序）：

```
lhs       rhs    support     confidence coverage   lift
{2069} => {748}  0.009677419 0.2307692  0.04193548 3.1103679
{1301} => {748}  0.009677419 0.2307692  0.04193548 3.1103679
{2050} => {748}  0.012903226 0.2105263  0.06129032 2.8375286
{4832} => {1924} 0.012903226 0.6666667  0.01935484 1.6939891
{948}  => {1924} 0.019354839 0.5454545  0.03548387 1.3859911
{3370} => {1924} 0.009677419 0.5000000  0.01935484 1.2704918
{1450} => {1924} 0.009677419 0.3750000  0.02580645 0.9528689
{4810} => {1924} 0.016129032 0.2941176  0.05483871 0.7473481
{2050} => {1924} 0.016129032 0.2631579  0.06129032 0.6686799
{1301} => {1924} 0.009677419 0.2307692  0.04193548 0.5863808
```

- 关联规则可视化及分析：





从上左图中可以清晰地发现，10 个关联规则全部指向商品 ID 为 748 和 1924 的两个明星产品。

	product_id	product_price	seller_id	sold_quantity	count	price_cluster	quantity_cluster	cluster
1178	1924	138.0	2378	59967.0	122	high-end	best-seller	best-seller ; high-end
1068	748	138.0	2378	1007.0	23	high-end	best-seller	best-seller ; high-end

从上右图中可以看到，【商品 1924】为这 10 个关联规则的中心。【商品 1924】是所有商品中历史销售量最高的商品，且远远大于第二名。这说明在商家 2378 购买电源适配器的消费者，大部分都会选择带上一个【商品 1924】。

3. 主要发现

3.1 电源适配器市场商品特点

在电源适配器市场中，大部分商品集中在中低端（价格<120 元），价格在 50 元左右。大多数商品总销量都在 100 以下，销量表现并不好，少数明星商品销量很高，最畅销的商品销量高达 59967。

3.2 消费者行为偏好

从购买频率上看，消费者喜欢购买“畅销品”和“优质品”，尽管它们种类不多。在价格上，偏好中低端价格的电源适配器商品。

从购买关联上看，消费者不喜欢跨价格类别购买商品。购买低价格产品的用户不太看重商品的历史销量，哪怕销量没那么高也会加入购物车。购买“高端优质品”的消费者很大程度上也会购买“高端畅销品”。

3.3 电源适配器市场竞争格局

从历史总销量来看，电源适配器市场的竞争格局集中，头部的两个商家（ID 分别为‘2378’和‘1575’）占据市场份额超过 70%。

3.4 明星店铺 2378 的特点

- 产品的价格类别集中，都是高端产品，价格在 120 元以上
- 产品种类较多，但是店铺 90%销售量主要靠 11 个“畅销高端品”
- 从购买关联上看，店铺依赖爆品 1924，可以认为爆品 1924 带动了店铺其他产品的销量

4. 研究结论及建议

本文有如下结论：

表 1 主要研究发现

	研究问题	研究发现	报告章节
1	市场产品特点是什么？哪些类型的产品最畅销？	产品特点：中低端，大部分商品销量一般，极个别销量很好 畅销产品类：中低端的“畅销品”和“优质品”	2.2.1；3.1
2	消费者喜欢同时购买哪些类型的产品？	消费者喜欢同时购买价格类别相同的产品。偏好低端产品的消费者不太看重销量；偏好高端产品的消费者会同时购买“畅销品”和“优质品”	2.2.2；3.2
3	哪些卖家表现最好，他们的特点是什么？	商家 2378 表现最好，其特点为商品价格类别集中，依赖爆品 1924 带动其他店铺中其他产品的销量	2.2.4；3.4

综上分析，本文给电源适配市场的商家提出如下建议以提高其销量：

- 将其商品集中在同一个价格类内（如集中打造高端商品，或集中打造性价比商品）
- 主打低价格商品的商家可以同时给顾客推荐销量不同的产品
- 主打高端商品的商家可以尝试打造个别爆品