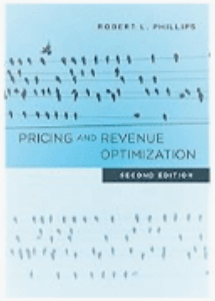
Market Basket Analysis

购物篮分析

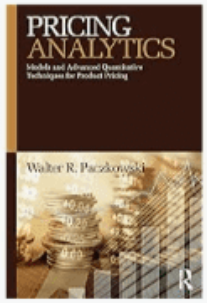
Have you heard about the story of
“diaper and beer”?

你听说过‘尿布和啤酒’的故事吗?

Frequently bought together



+



This item: Pricing and Revenue Optimization: Second Edition

$115.00

Pricing Analytics: Models and Advanced Quantitative Techniques for Product Pricing

$59.54

Total price: $174.54

Add both to Cart

One of these items ships sooner than the other.

Show details

天猫 darlie黑人旗舰店

黑人星耀白雪绒花牙膏去黄去牙渍亮白口腔清洁清新口气男女生含氟 ¥89.90 x1

净含量:星耀白牙膏120g*4支-买就送牙膏40g*3（不倍增）；

七天无理由退换

购买数量 − 1 +

配送方式 普通配送 快递 免邮 >

运费险 卖家赠送，退换货可赔 >

店铺优惠 省20元:组合优惠 -¥20.00 >

开具发票 本次不开具发票 >

订单备注 选填,请先和商家协商一致

共1件 小计: ¥69.90

顺手买一件

黑人密护龈软毛牙刷情侣小头牙[illegible]清洁工具女男士专用口腔清洁成[illegible]

颜色分类:密护龈2支装（颜色随机，不参与买赠以及满减活动）；

现价 ¥9.90 价格 ¥19.90

该笔使用安全免密支付，提交订单直接支付

共1件，合计: ¥69.90 提交订单

This question is also relevant for financial practitioners. For instance, there are about 2,500 stocks traded in the Hong Kong Stock Exchange, and an investor typically holds multiple stocks. By using similar analysis, we can see which stocks investors tend to hold together, and you can make recommendations to your clients accordingly.

这个问题对金融从业者也很重要。例如，在香港证券交易所有大约2,500只股票在交易，投资者通常持有多只股票。通过类似的分析，我们可以看到投资者倾向于同时持有哪些股票，然后您可以相应地向您的客户提供建议。

We examine a strategy to extract insight from transactions and cooccurrence data: association rule mining. Association rule analysis attempts to find sets of informative patterns from large, sparse data sets.

Which products do consumers purchase together?

Which stocks do investors invest together?

Which services do clients use together?

我们研究一种从交易和共现数据中提取见解的策略：关联规则挖掘。关联规则分析试图从大型稀疏数据集中找到一组信息丰富的模式。

消费者一起购买哪些产品？

投资者一起投资哪些股票？

客户一起使用哪些服务？

The Basic Idea

Suppose that 2% of your shoppers buy diapers and 5% of them buy beer in your supermarket.

Now, let us focus on those who buy diapers. Among these shoppers, if 5% of them also buy beer, you can claim that diaper-buyers do not like beer more or less than others do, and there is no specific relationship between diaper and beer. However, if 25% of them also buy beer, it is quite different than the base rate and is evidence of an association.

基本理念

假设你的顾客中有2%购买尿布，5%购买啤酒。

现在，让我们关注那些购买尿布的顾客。在这些购买尿布的顾客中，如果有5%的人也购买啤酒，你可以断定购买尿布的人对啤酒的需求不会更多或更少，尿布和啤酒之间没有特定的关系。然而，如果有25%的人也购买啤酒，那就与基础率相去甚远，这是关联的证据。

Background

A transaction, or a market basket, is the set of things that are purchases at one occasion. For each, {beer, diaper, chocolate} is a transaction of a consumer.

A rule expresses the incidence across transactions of one set of items as a condition of another set of items. It can be something like {diaper}->{beer}, but can also be like {potato, chocolate}->{beer, soda, water}.

背景

交易或购物篮是一次购买的物品集合。例如，{啤酒，尿布，巧克力} 是一个消费者的交易。

规则表达了一组物品在交易中的发生与另一组物品出现的关联。它可以是像{尿布}->{啤酒}这样的规则，也可以是像{土豆，巧克力}->{啤酒，苏打水，水}这样的规则。

Metrics

The support for a set of items is the proportion of all transactions that contain the set. For example, if {pizza, soda} appears in 10 out of 200 transactions, then

$$\text{support}(\text{pizza}, \text{soda}) = \frac{10}{200} = 0.05.$$

度量指标

一组物品的支持度 (support) 是包含该组物品的所有交易的比例。例如，如果{比萨，苏打水}在200次交易中出现了10次，则

$$\text{support}(\text{pizza}, \text{soda}) = \frac{10}{200} = 0.05.$$

Metrics

Confidence is the support for the cooccurrence of all items in a rule, conditional on the support for the left hand set alone.

$$\text{confidence}(X \to Y) = \frac{\text{support}(X \text{and } Y)}{\text{support X}}.$$

Equivalently, confidence$(X \to Y)$ measures how likely a consumer purchases Y given that the consumer already purchases X.

度量指标

置信度(Confidence) 被定义为当左边产品组合被购买的情况下，右边产品组合被购买的概率.

$$\text{confidence}(X \to Y) = \frac{\text{support}(X \text{and } Y)}{\text{support X}}.$$

换句话说, $\text{confidence}(X \to Y)$ 表示消费者已经买了 X 的情况下会同时购买 Y 的概率。

Metrics

Note that confidence$(X \to Y)$ is not always equal to confidence$(Y \to X)$, for instance:

$\text{confidence}(\text{MBA} \to \text{Bachelor}) = 1$: If a person has an MBA degree, he/she must also have a bachelor degree.

$\text{confidence}(\text{Bachelor} \to \text{MBA}) = 0.05$: If a person has a bachelor degree, with probability 5% he or she also has an MBA degree.

度量指标

需要注意的是， confidence($X \to Y$) 并不总是等同于 confidence($Y \to X$), 举例来说:

confidence(MBA $\to$ Bachelor) $= 1$: 如果一个人具有MBA学位，那么这个人一定具有本科学位

confidence(Bachelor $\to$ MBA) $= 0.05$: 如果一个人具有本科学位，那么这个人有 5% 的几率具有 MBA 学位。

Metrics

A more important measure, lift, is the support of a set conditional on the joint support of each element:

$$\text{lift}(X \to Y) = \frac{\text{support}(X \text{and } Y)}{\text{support}(X) \times \text{support}(Y)}.$$

When lift is greater than 1, it means the two items are likely to occur together. The larger lift is, the stronger the connection between the items.

度量指标

一个更重要的指标, 提升度(lift), 表示两组产品同时购买和分别购买之间的关系:

$$\text{lift}(X \to Y) = \frac{\text{support}(X \text{and } Y)}{\text{support}(X) \times \text{support}(Y)}.$$

当 lift 大于 1 时，表示这两组产品更可能被同时购买。而这个概率越大，则表示这两组产品越有机会被同时购买。

First, we load data of consumer purchase information.

```
1 library(arules)
2 library(arulesViz)
3 mydata = readLines("https://ximarketing.github.io/data/basket.txt")
4 head(mydata)
```

```
[1] "0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 "
[2] "30 31 32 "
[3] "33 34 35 "
[4] "36 37 38 39 40 41 42 43 44 45 46 "
[5] "38 39 47 48 "
[6] "38 39 48 49 50 51 52 53 54 55 56 57 58 "
```

The second consumer has bought products number 30, 31, and 32 during at one occasion.

我们首先载入数据

```
1 library(arules)
2 library(arulesViz)
3 mydata = readLines("https://ximarketing.github.io/data/basket.txt")
4 head(mydata)
```

```
[1] "0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 "
[2] "30 31 32 "
[3] "33 34 35 "
[4] "36 37 38 39 40 41 42 43 44 45 46 "
[5] "38 39 47 48 "
[6] "38 39 48 49 50 51 52 53 54 55 56 57 58 "
```

第二个消费者一次同时购买了编号为 30, 31, 和 32 的产品

```
1 mydata = strsplit(mydata, " ")
2 transactions <- as(mydata, "transactions")
3 summary(transactions)
```

Next, we create transaction records from the data, which can be used for further analysis.

```
most frequent items:
     39       48       38       32       41 (Other)
  50675    42135    15596    15167    14945   770058
```

These are the most popular items in the transaction records.

```
1 mydata = strsplit(mydata, " ")
2 transactions <- as(mydata, "transactions")
3 summary(transactions)
```

接下来，我们从数据中创建交易记录，这可以用于进一步分析。

```
most frequent items:
      39       48       38       32       41 (Other)
   50675    42135    15596    15167    14945   770058
```

这些是交易记录中最受欢迎的物品。

```
rules <- apriori(transactions,
                 parameter= list(supp=0.001, conf=0.4))
inspect(sort(rules, by="lift"))
```

This line allows us to create the association rules $\{A\} \rightarrow \{B\}$, with two restrictions: (1) The support should be at least 0.001, and the confidence should be at least 0.4. We then sort the rules by their lift and show the results.

```
rules <- apriori(transactions,
                 parameter= list(supp=0.001, conf=0.4))
inspect(sort(rules, by="lift"))
```

这一行允许我们创建关联规则 $\{A\} \rightarrow \{B\}$，并有两个限制条件：(1) 支持度 support 至少为0.001，置信度 confidence 至少为0.4。然后，我们根据提升度 lift 对规则进行排序，并展示结果。

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{696}	=>	{699}	0.001032191	0.5833333	0.001769470	338.3410	91
[2]	{699}	=>	{696}	0.001032191	0.5986842	0.001724099	338.3410	91
[3]	{1818, 3311, 795}	=>	{1819}	0.001088905	0.9056604	0.001202332	318.1069	96
[4]	{3402}	=>	{3535}	0.001417844	0.7062147	0.002007668	305.2024	125
[5]	{3535}	=>	{3402}	0.001417844	0.6127451	0.002313922	305.2024	125
[6]	{1818, 1819, 795}	=>	{3311}	0.001088905	0.8275862	0.001315760	302.7455	96
[7]	{1819, 3311, 795}	=>	{1818}	0.001088905	0.7741935	0.001406502	302.0108	96
[8]	{3311, 795}	=>	{1819}	0.001406502	0.8435374	0.001667385	296.2866	124
[9]	{1818, 1819, 3311}	=>	{795}	0.001088905	0.8421053	0.001293074	295.7836	96
[10]	{3537, 39}	=>	{3535}	0.001043533	0.6764706	0.001542615	292.3480	92

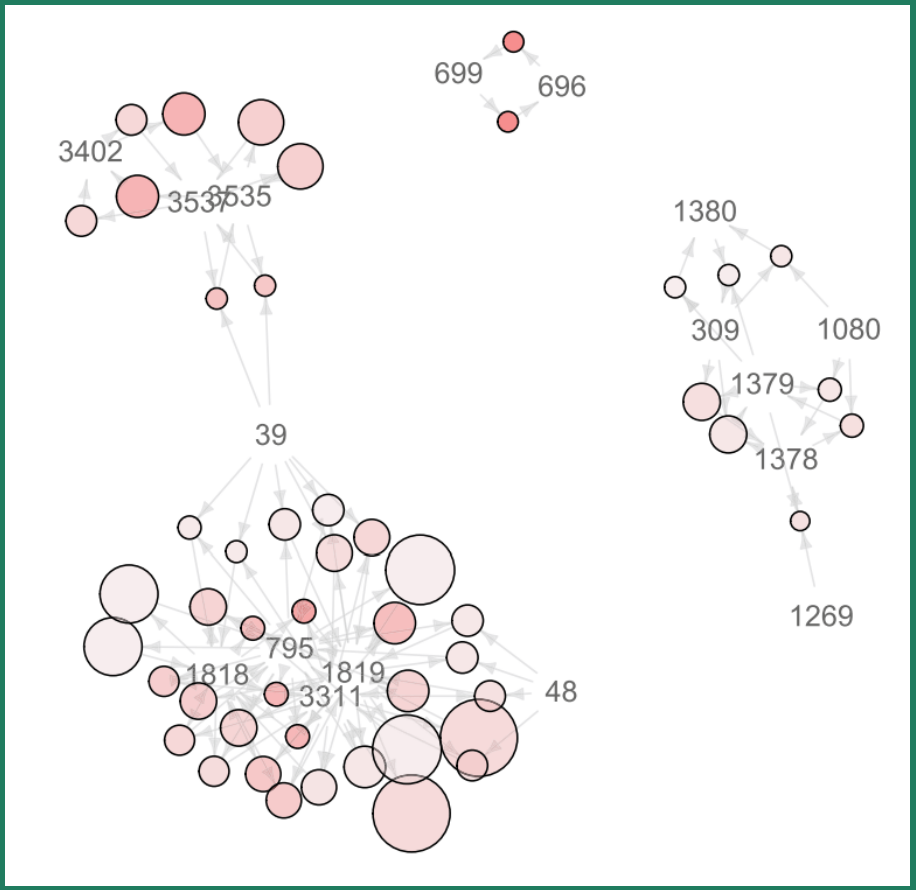
These are the top 10 rules that we detected, and you can use the result to make recommendations to your consumers. For example, if one consumer buys item 696, you can ask the consumer "do you want to buy item 699 with it?"

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{696}	=>	{699}	0.001032191	0.5833333	0.001769470	338.3410	91
[2]	{699}	=>	{696}	0.001032191	0.5986842	0.001724099	338.3410	91
[3]	{1818, 3311, 795}	=>	{1819}	0.001088905	0.9056604	0.001202332	318.1069	96
[4]	{3402}	=>	{3535}	0.001417844	0.7062147	0.002007668	305.2024	125
[5]	{3535}	=>	{3402}	0.001417844	0.6127451	0.002313922	305.2024	125
[6]	{1818, 1819, 795}	=>	{3311}	0.001088905	0.8275862	0.001315760	302.7455	96
[7]	{1819, 3311, 795}	=>	{1818}	0.001088905	0.7741935	0.001406502	302.0108	96
[8]	{3311, 795}	=>	{1819}	0.001406502	0.8435374	0.001667385	296.2866	124
[9]	{1818, 1819, 3311}	=>	{795}	0.001088905	0.8421053	0.001293074	295.7836	96
[10]	{3537, 39}	=>	{3535}	0.001043533	0.6764706	0.001542615	292.3480	92

这些是我们检测到的前10条规则，您可以利用这些结果向您的消费者提供建议。例如，如果一个消费者购买了商品696，您可以询问消费者：“您是否想一起购买商品699呢？”

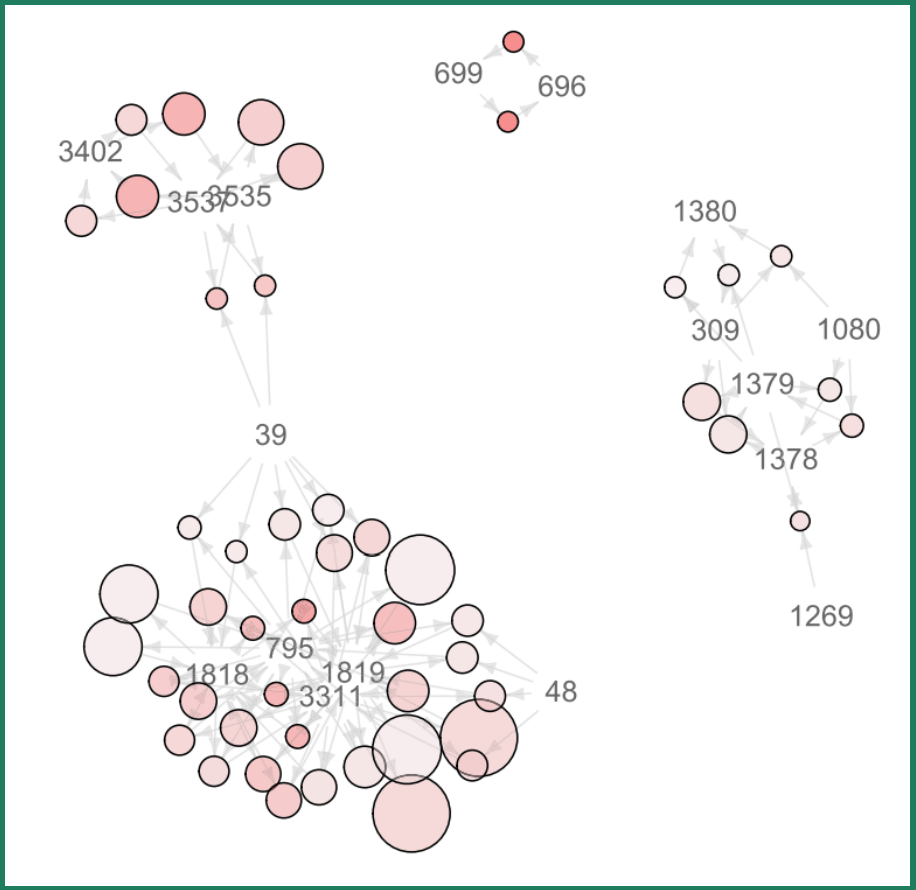
```
1 plot(rules , method="graph", control= list(type="items"))
```

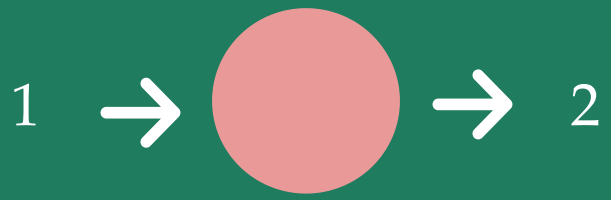
We can further visualize the rules we have detected. You will get something like this (it varies with different for the system):



```
1 plot(rules , method="graph", control= list(type="items"))
```

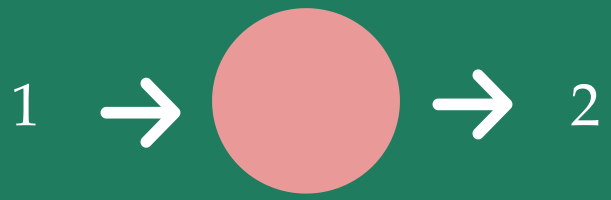
我们可以进一步将我们检测到的规则进行可视化展示。您将会看到类似以下内容的展示（实际展示会因系统不同而有所变化）：





In the above visualization, each circle represents a rule. The inbound arrow captures the items on the left-hand side of the rule, and the outbound arrow captures the items on the right-hand side of the rule. Here, we have a rule {1}->{2}.

The size (area) of the circle represents the rule's support, and shade represents lift (darker indicates higher lift).



在上述可视化中，每个圆代表一条规则。指向圆的箭头代表了规则左侧的项目，指出圆的箭头代表了规则右侧的项目。这里我们有一条规则{1}->{2}。

圆的大小（面积）代表规则的支持度 support，颜色深浅代表提升度 lift（颜色越深表示 lift 越高）。

The complete code is here:

```
library(arules)
library(arulesViz)
mydata = readLines("https://ximarketing.github.io/data/basket.txt")
mydata = strsplit(mydata, " ")
transactions <- as(mydata, "transactions")
rules <- apriori(transactions,
    parameter= list(supp=0.001, conf=0.4))
inspect(sort(rules, by="lift"))
plot(rules , method="graph", control= list(type="items"))
```

A Stock Example 一个股票的例子

You are a stock broker. You would like to recommend stocks to investors. You can analyze the stock holdings of individual investors and find some rules. For example, you may reach the conclusion that "if a person invests in CCB, the person may also be interested in ICBC."

你是一名股票经纪人，希望向投资者推荐股票。你可以分析个人投资者的股票持有情况，并找出一些规律。例如，你可能得出结论："如果一个人投资了中国建设银行，那么这个人可能也对中国工商银行感兴趣。"

A Stock Example 一个股票的例子

I don’t have the information about individual stock holdings, but I can access the holdings of mutual funds.

我没有关于个人股票持有情况的信息，但我可以访问共同基金的持仓情况。

You can access the funds information here.

你可以在这里查询基金信息

（基金代码）基金名称	（基金代码）基金名称	（基金代码）基金名称
（000001）华夏成长混合 | 基金吧 | 档案	（000003）中海可转债债券A | 基金吧 | 档案	（000004）中海可转债债券C | 基金吧 | 档案
（000005）嘉实增强信用定期债券 | 基金吧 | 档案	（000006）西部利得量化成长混合A | 基金吧 | 档案	（000008）嘉实中证500ETF联接A | 基金吧 | 档案
（000009）易方达天天理财货币A | 基金吧 | 档案	（000010）易方达天天理财货币B | 基金吧 | 档案	（000011）华夏大盘精选混合A | 基金吧 | 档案
（000013）易方达天天理财货币R | 基金吧 | 档案	（000014）华夏聚利债券A | 基金吧 | 档案	（000015）华夏纯债债券A | 基金吧 | 档案
（000016）华夏纯债债券C | 基金吧 | 档案	（000017）财通可持续混合 | 基金吧 | 档案	（000020）景顺长城品质投资混合A | 基金吧 | 档案
（000021）华夏优势增长混合 | 基金吧 | 档案	（000024）大摩双利增强债券A | 基金吧 | 档案	（000025）大摩双利增强债券C | 基金吧 | 档案
（000028）华富安鑫债券A | 基金吧 | 档案	（000029）富国宏观策略灵活配置混合A | 基金吧 | 档案	（000030）长城核心优选混合A | 基金吧 | 档案
（000031）华夏复兴混合A | 基金吧 | 档案	（000032）易方达信用债债券A | 基金吧 | 档案	（000033）易方达信用债债券C | 基金吧 | 档案
（000037）广发景宁债券A | 基金吧 | 档案	（000039）农银高增长混合 | 基金吧 | 档案	（000041）华夏全球股票(QDII)(| 基金吧 | 档案

You can further check the holdings of individual funds.

你可以进一步查询每个基金的持仓

股票持仓 债券持仓 更多 >

股票名称	持仓占比	涨跌幅	相关资讯
航天电器	5.33%	-2.93%	股吧
中航高科	3.83%	-1.90%	股吧
菲利华	3.25%	-3.36%	股吧
立讯精密	2.30%	-1.39%	股吧
三环集团	2.15%	-1.15%	股吧
科大讯飞	1.87%	-2.10%	股吧
北方华创	1.84%	-2.37%	股吧
海康威视	1.70%	-1.76%	股吧
海光信息	1.65%	-3.86%	股吧
恒瑞医药	1.42%	-3.82%	股吧

```r
library(arules)
library(arulesViz)
options(encoding = "GB18030")
mydata <-
readLines("https://ximarketing.github.io/data/fund_holdings.txt",
encoding = "GB18030")
head(mydata)
mydata = strsplit(mydata, " ")
transactions <- as(mydata, "transactions")
rules <- apriori(transactions,
                 parameter= list(supp=0.01, conf=0.1))
inspect(sort(rules, by="lift"))
plot(rules , method="graph", control= list(type="items"))
```


泸州老窖
亿纬锂能
五粮液
阳光电源
山西汾酒
立讯精密
沪电股份
贵州茅台
宁德时代
新易盛
中际旭创
招商银行
中芯国际
长江电力
寒武纪-U
中国平安
美的集团
兴业银行
海光信息
工商银行
北方华创
恒瑞医药
陕西煤业
腾讯控股
建设银行
中国银行
小米集团-W
阿里巴巴-W
天孚通信
农业银行
中信证券
中微公司
科伦药业
中国神华
美团-W
lift
20.0
17.5
15.0
12.5
10.0
7.5
support
0.010
0.015
0.020
0.025
0.030
0.035

	lhs				rhs	support	confidence	coverage
[1]	{中国平安, 招商银行, 贵州茅台}			=>	{兴业银行}	0.01065217	0.4827586	0.02206522
[2]	{中国平安, 招商银行}			=>	{兴业银行}	0.01163043	0.3780919	0.03076087
[3]	{兴业银行, 招商银行, 贵州茅台}			=>	{中国平安}	0.01065217	0.9158879	0.01163043
[4]	{宁德时代, 阳光电源}			=>	{亿纬锂能}	0.01119565	0.3259494	0.03434783
[5]	{中国平安, 贵州茅台}			=>	{兴业银行}	0.01195652	0.3606557	0.03315217
[6]	{中际旭创, 寒武纪-U}			=>	{新易盛}	0.01032609	0.8260870	0.01250000
[7]	{兴业银行, 贵州茅台}			=>	{中国平安}	0.01195652	0.8396947	0.01423913
[8]	{宁德时代, 招商银行, 贵州茅台, 长江电力}			=>	{中国平安}	0.01119565	0.8240000	0.01358696
[9]	{宁德时代, 美的集团, 贵州茅台, 长江电力}			=>	{中国平安}	0.01195652	0.8088235	0.01478261
[10]	{招商银行, 贵州茅台}			=>	{兴业银行}	0.01163043	0.3203593	0.03630435
[11]	{兴业银行, 招商银行}			=>	{中国平安}	0.01163043	0.7867647	0.01478261
[12]	{中国银行}			=>	{农业银行}	0.01130435	0.5333333	0.02119565
[13]	{农业银行}			=>	{中国银行}	0.01130435	0.2708333	0.04173913
[14]	{亿纬锂能, 宁德时代}			=>	{阳光电源}	0.01119565	0.5919540	0.01891304
[15]	{宁德时代, 招商银行, 美的集团, 贵州茅台}			=>	{中国平安}	0.01336957	0.7592593	0.01760870
[16]	{宁德时代, 贵州茅台, 长江电力}			=>	{中国平安}	0.01369565	0.7500000	0.01826087
[17]	{宁德时代, 招商银行, 长江电力}			=>	{中国平安}	0.01184783	0.7364865	0.01608696
[18]	{天孚通信}			=>	{中际旭创}	0.01173913	0.7105263	0.01652174
[19]	{中际旭创}			=>	{天孚通信}	0.01173913	0.2022472	0.05804348

If an investor holds CCB stock, what else will she/he consider?

如果一个投资者持有建设银行的股票，他/她还会考虑什么？

```
rules <- apriori(transactions, parameter = list(supp = 0.001,
conf = 0.01))
A <- "建设银行"
rules_A <- subset(rules, lhs %pin% A)
top_rules <- head(sort(rules_A, by = "confidence"), 20)
inspect(top_rules)
```

[1]	{电投能源，建设银行}	=>	{工商银行}	0.001521739	1	0.001521739 16.576577 14
[2]	{建设银行，中国联通}	=>	{中国电信}	0.001304348	1	0.001304348 44.878049 12
[3]	{国投电力，建设银行}	=>	{长江电力}	0.001304348	1	0.001304348 11.825193 12
[4]	{汇丰控股，建设银行，美团-W}	=>	{腾讯控股}	0.001304348	1	0.001304348 6.637807 12
[5]	{电投能源，建设银行，中国神华}	=>	{中国银行}	0.001086957	1	0.001086957 47.179487 10
[6]	{电投能源，建设银行，中国银行}	=>	{工商银行}	0.001304348	1	0.001304348 16.576577 12
[7]	{电投能源，建设银行，中国银行}	=>	{长江电力}	0.001304348	1	0.001304348 11.825193 12
[8]	{电投能源，建设银行，长江电力}	=>	{中国银行}	0.001304348	1	0.001304348 47.179487 12
[9]	{电投能源，建设银行，农业银行}	=>	{工商银行}	0.001304348	1	0.001304348 16.576577 12
[10]	{电投能源，建设银行，中国神华}	=>	{工商银行}	0.001086957	1	0.001086957 16.576577 10
[11]	{电投能源，建设银行，中国神华}	=>	{长江电力}	0.001086957	1	0.001086957 11.825193 10
[12]	{电投能源，建设银行，长江电力}	=>	{工商银行}	0.001304348	1	0.001304348 16.576577 12
[13]	{建设银行，交通银行，中国银行}	=>	{农业银行}	0.001413043	1	0.001413043 23.958333 13
[14]	{建设银行，中国海油，中国神华}	=>	{工商银行}	0.001630435	1	0.001630435 16.576577 15
[15]	{阿里巴巴-W，建设银行，小米集团-W}	=>	{美团-W}	0.001086957	1	0.001086957 19.870410 10
[16]	{阿里巴巴-W，建设银行，美团-W}	=>	{腾讯控股}	0.001521739	1	0.001521739 6.637807 14
[17]	{阿里巴巴-W，建设银行，小米集团-W}	=>	{腾讯控股}	0.001086957	1	0.001086957 6.637807 10
[18]	{阿里巴巴-W，建设银行，招商银行}	=>	{腾讯控股}	0.001413043	1	0.001413043 6.637807 13
[19]	{阿里巴巴-W，建设银行，宁德时代}	=>	{腾讯控股}	0.001521739	1	0.001521739 6.637807 14
[20]	{格力电器，建设银行，宁德时代}	=>	{美的集团}	0.001195652	1	0.001195652 5.361305 11

Who would you recommend CCB stock to?

你想把建设银行股票推荐给谁

```
rules <- apriori(transactions, parameter = list(supp = 0.001,
conf = 0.01))
A <- "建设银行"
rules_A <- subset(rules, rhs %pin% A)
top_rules <- head(sort(rules_A, by = "confidence"), 20)
inspect(top_rules)
```

```
[1]  {电投能源，中国银行}                                  => {建设银行} 0.001304348 0.7500000  0.001739130 22.84768 12
[2]  {电投能源，工商银行，中国银行}                        => {建设银行} 0.001304348 0.7500000  0.001739130 22.84768 12
[3]  {电投能源，长江电力，中国银行}                        => {建设银行} 0.001304348 0.7500000  0.001739130 22.84768 12
[4]  {电投能源，工商银行，长江电力}                        => {建设银行} 0.001304348 0.7500000  0.001739130 22.84768 12
[5]  {电投能源，工商银行，长江电力，中国银行}              => {建设银行} 0.001304348 0.7500000  0.001739130 22.84768 12
[6]  {电投能源，工商银行}                                  => {建设银行} 0.001521739 0.7368421  0.002065217 22.44685 14
[7]  {农业银行，中煤能源}                                  => {建设银行} 0.001195652 0.7333333  0.001630435 22.33996 11
[8]  {电投能源，农业银行，中国银行}                        => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
[9]  {电投能源，中国神华，中国银行}                        => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
[10] {电投能源，农业银行，长江电力}                        => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
[11] {电投能源，工商银行，中国神华}                        => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
[12] {电投能源，长江电力，中国神华}                        => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
[13] {电投能源，工商银行，农业银行，中国银行}              => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
[14] {电投能源，农业银行，长江电力，中国银行}              => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
[15] {电投能源，工商银行，中国神华，中国银行}              => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
[16] {电投能源，长江电力，中国神华，中国银行}              => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
[17] {电投能源，工商银行，农业银行，长江电力}              => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
[18] {电投能源，工商银行，长江电力，中国神华}              => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
[19] {电投能源，工商银行，农业银行，长江电力，中国银行} => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
[20] {电投能源，工商银行，长江电力，中国神华，中国银行} => {建设银行} 0.001086957 0.7142857  0.001521739 21.75970 10
```

Who would you recommend CCB stock to?

你想把建设银行股票推荐给谁

```
rules <- apriori(transactions, parameter = list(supp = 0.001, conf = 0.01))
A <- "建设银行"
rules_with_one_item <- subset(rules, ((size(lhs) == 1) & (rhs %pin% A)))
top_rules <- head(sort(rules_with_one_item, by = "confidence"), 20)
inspect(top_rules)
```

[1]	{汇丰控股}	=>	{建设银行}	0.002173913	0.3636364	0.005978261	11.077664	20
[2]	{安迪苏}	=>	{建设银行}	0.001086957	0.3225806	0.003369565	9.826960	10
[3]	{中国银行}	=>	{建设银行}	0.006413043	0.3025641	0.021195652	9.217185	59
[4]	{景津装备}	=>	{建设银行}	0.001413043	0.2708333	0.005217391	8.250552	13
[5]	{中国石油股份}	=>	{建设银行}	0.001086957	0.2702703	0.004021739	8.233399	10
[6]	{工商银行}	=>	{建设银行}	0.015652174	0.2594595	0.060326087	7.904063	144
[7]	{中信银行}	=>	{建设银行}	0.001739130	0.2580645	0.006739130	7.861568	16
[8]	{农业银行}	=>	{建设银行}	0.010326087	0.2473958	0.041739130	7.536562	95
[9]	{光大银行}	=>	{建设银行}	0.001304348	0.2400000	0.005434783	7.311258	12
[10]	{中国石油}	=>	{建设银行}	0.003586957	0.2200000	0.016304348	6.701987	33
[11]	{华电国际}	=>	{建设银行}	0.001195652	0.2037037	0.005869565	6.205543	11
[12]	{中国建筑}	=>	{建设银行}	0.004673913	0.1945701	0.024021739	5.927302	43
[13]	{重庆农村商业银行}	=>	{建设银行}	0.001304348	0.1690141	0.007717391	5.148773	12
[14]	{交通银行}	=>	{建设银行}	0.002500000	0.1554054	0.016086957	4.734204	23
[15]	{中国神华}	=>	{建设银行}	0.008804348	0.1525424	0.057717391	4.646986	81
[16]	{邮储银行}	=>	{建设银行}	0.001847826	0.1504425	0.012282609	4.583016	17
[17]	{中国电信}	=>	{建设银行}	0.003043478	0.1365854	0.022282609	4.160879	28
[18]	{粤高速Ａ}	=>	{建设银行}	0.001195652	0.1358025	0.008804348	4.137029	11
[19]	{电投能源}	=>	{建设银行}	0.001521739	0.1296296	0.011739130	3.948982	14
[20]	{京沪高铁}	=>	{建设银行}	0.001195652	0.1294118	0.009239130	3.942345	11

Conditional Logit Model and Conjoint Analysis

条件Logit模型和联合分析

In multinomial logit model, a person chooses among a few alternatives. The decision hinges on the decision maker's personal features, not the features of the alternatives. In our previous example, the route decision hinges on features such as distance, age, which are constant across all alternatives.

In conditional logit model, a person chooses among a few alternatives. The decision hinges on the alternatives' features, not the feature of the individuals.

在 multinomial logit model 模型, 一个人从几个选项中做出选择。 这个选择取决于这个人的个人特征而不是这些选项的特征，例如，选择基于这个人的年龄，性别等个人特征。

在 conditional logit model 模型, 一个人从几个选项中做出选择。这个选择取决于这个选项的特征而不是这个人的特征。例如，选择基于这个选项的价格，质量，颜色等。

Example:

Consumers choose among three computers, A, B, and C.

1. If the choices are based on consumers' age, gender, education etc, then we use the multinomial logit model.
2. If the choices are based on the price, quality of the computers, then we use the conditional logit model.

举例:

消费者从三个电脑品牌中选择一个, A, B, 和 C.

1. 如果选择是基于消费者的年龄，性别，职业等信息，那么我们选择的模型是 multinomial logit model.
2. 如果选择是基于每个电脑的价格，质量，服务等，那么我们选择的模型是 conditional logit model.

```r
install.packages("survival")
library(survival)
library(stargazer)
mydata = read.csv("https://ximarketing.github.io/data/conjoint.csv")
head(mydata)
```

id	interest	downpayment	rebate	speed	choice
1	3.75	40	0.15	0.5	0
1	4.00	25	0.15	1.0	0
1	3.75	25	0.00	1.0	1
2	3.50	20	0.10	0.5	1
2	3.75	25	0.30	1.5	0
2	3.75	20	0.30	1.0	0

id	interest	downpayment	rebate	speed	choice
1	3.75	40	0.15	0.5	0
1	4.00	25	0.15	1.0	0
1	3.75	25	0.00	1.0	1
2	3.50	20	0.10	0.5	1
2	3.75	25	0.30	1.5	0
2	3.75	20	0.30	1.0	0

Consumer 1 (id = 1) chooses between three offers:

Interest Rate	Down Payment	Rebate	Speed (Months)	Choice
3.75%	40%	0.15%	0.5	NO
4.00%	25%	0.15%	1.0	NO
3.75%	25%	0%	1.0	YES

id	interest	downpayment	rebate	speed	choice
1	3.75	40	0.15	0.5	0
1	4.00	25	0.15	1.0	0
1	3.75	25	0.00	1.0	1
2	3.50	20	0.10	0.5	1
2	3.75	25	0.30	1.5	0
2	3.75	20	0.30	1.0	0

用户1 (id = 1) 从下面三个按揭计划中做出选择:

按揭利率	首付	回赠	审批时间（月度）	选择
3.75%	40%	0.15%	0.5	NO
4.00%	25%	0.15%	1.0	NO
3.75%	25%	0%	1.0	YES

```
result<-clogit(choice ~ interest +  downpayment + rebate
               + speed + strata(id), data=mydata)
summary(result)
```

```
                coef exp(coef)  se(coef)      z Pr(>|z|)
interest    -1.185055  0.305729  0.097289 -12.181  < 2e-16 ***
downpayment -0.052922  0.948454  0.002336 -22.652  < 2e-16 ***
rebate       0.177522  1.194254  0.149303   1.189  0.23444
speed       -0.117274  0.889341  0.039587  -2.962  0.00305 **
```

```
stargazer(result, type="html", out="result.html")
```

	Dependent variable:
	choice
interest	-1.185***
	(0.097)
downpayment	-0.053***
	(0.002)
rebate	0.178
	(0.149)
speed	-0.117***
	(0.040)
Observations	18,000
R^2	0.039
Max. Possible R^2	0.519
Log Likelihood	-6,237.584
Wald Test	643.940*** (df = 4)
LR Test	708.180*** (df = 4)
Score (Logrank) Test	679.185*** (df = 4)
Note:	*p<0.1; **p<0.05; ***p<0.01

When interest rate increases, the user is less likely to choose the plan; when down payment increases, the user is less likely to choose the plan; when approval takes longer time, the user is less likely to choose the plan.

当利率增加，首付变高，或者审批时间变长的情况下，用户选择按揭的概率降低。

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
interest	-1.185055	0.305729	0.097289	-12.181	< 2e-16	***
downpayment	-0.052922	0.948454	0.002336	-22.652	< 2e-16	***
rebate	0.177522	1.194254	0.149303	1.189	0.23444	
speed	-0.117274	0.889341	0.039587	-2.962	0.00305	**

The coefficient for interest is -1.185 and the coefficient for speed is -0.1172. Because 0.1172/1.185=0.098, it suggests that a 1 month increase in approval time is equivalent to a 0.098% decrease in interest rate.

利率的系数是-1.185而审批时间系数是-0.1172. 因为0.1172/1.185=0.098，这说明审批时间增加一个月的代价相当于利率上涨0.098%带来的代价。换句话说，一个月的审批时间对于消费者的价值是0.098%的利率。

Conjoint Analysis

Conjoint analysis is another useful tool for setting your prices, especially for existing products that consumers are familiar with. Let us image that you are determining the best interest to offer to clients, where each plan has a few attributes: interest rate, down payment, rebate, and review time.

联合分析

联合分析是设置价格的一个有用工具，尤其适用于消费者熟悉的现有产品。让我们设想你正在确定向客户提供的最佳利率，其中每个计划都有几个属性：利率、首付款、现金回赠和审核时间。

Conjoint Analysis

You then create different combinations and let consumers choose from the alternatives like this:

Interest Rate	Down Payment	Rebate	Review Time
3.75%	40%	0.15%	0.5
4.00%	25%	0.15%	1.0
3.75%	25%	0%	1.0

And for different clients, you make different choice sets and let them make the choice.

联合分析

然后，您创建不同的组合，让消费者从类似以下的备选方案中进行选择：

贷款利率	首付	现金回赠	审批时间
3.75%	40%	0.15%	0.5
4.00%	25%	0.15%	1.0
3.75%	25%	0%	1.0

对于不同的消费者，您制定不同的选择集，并让他们进行选择。

Conjoint Analysis

For example, you survey 6,000 consumers, and each consumer chooses among 3 alternatives. Then you plug the data into your conditional logistic model, and get results like this:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
interest	-1.185055	0.305729	0.097289	-12.181	< 2e-16	***
downpayment	-0.052922	0.948454	0.002336	-22.652	< 2e-16	***
rebate	0.177522	1.194254	0.149303	1.189	0.23444	
speed	-0.117274	0.889341	0.039587	-2.962	0.00305	**

联合分析

例如，您对6,000名消费者进行调查，每名消费者在3个备选方案中进行选择。然后，您将数据输入到 Conditional Logit 模型中，并获得以下结果：

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
interest	-1.185055	0.305729	0.097289	-12.181	< 2e-16	***
downpayment	-0.052922	0.948454	0.002336	-22.652	< 2e-16	***
rebate	0.177522	1.194254	0.149303	1.189	0.23444	
speed	-0.117274	0.889341	0.039587	-2.962	0.00305	**

Conjoint Analysis

Then, you can answer questions like this:

Given the offers of my competitors, if my interest rate decreases by 1%, how would my market share change?

•

联合分析

然后，您可以回答以下类似的问题：

如果我的利率降低1%，鉴于我的竞争对手的按揭计划，我的市场份额会如何变化？

```r
library(survival)
library(stargazer)
mydata = read.csv("https://ximarketing.github.io/data/conjoint.csv")
head(mydata)
result<-clogit(choice ~ interest +  downpayment +   rebate
               + speed + strata(id), data=mydata)
coef_interest <- coef(result)["interest"]
coef_downpayment <- coef(result)["downpayment"]
coef_rebate <- coef(result)["rebate"]
coef_speed <- coef(result)["speed"]

interest1 <- 3.85; downpayment1 <- 30; rebate1 <- 0.1; speed1 <- 1
interest2 <- 4.25; downpayment2 <- 25; rebate2 <- 0.25; speed2 <- 0.5

d1 <- exp(interest1 * coef_interest + downpayment1 * coef_downpayment +
            rebate1 * coef_rebate + speed1 * coef_speed)
d2 <- exp(interest2 * coef_interest + downpayment2 * coef_downpayment +
            rebate2 * coef_rebate + speed2 * coef_speed)

s1 <- d1/(d1+d2)
s2 <- d2/(d1+d2)
print(c(s1, s2))
```

Suppose that you are designing the first plan.

If you keep interest rate to 3.85%, your market share is 53.1%.

If you raise interest rate to 4.85%, your market share drops to 25.7%.

If you raise interest rate to 5.85%, your market share drops to 9.5%.

If you cut interest rate to 2.85%, your market share increases to 78.7%.

You can choose the interest that balances your margin and market share!

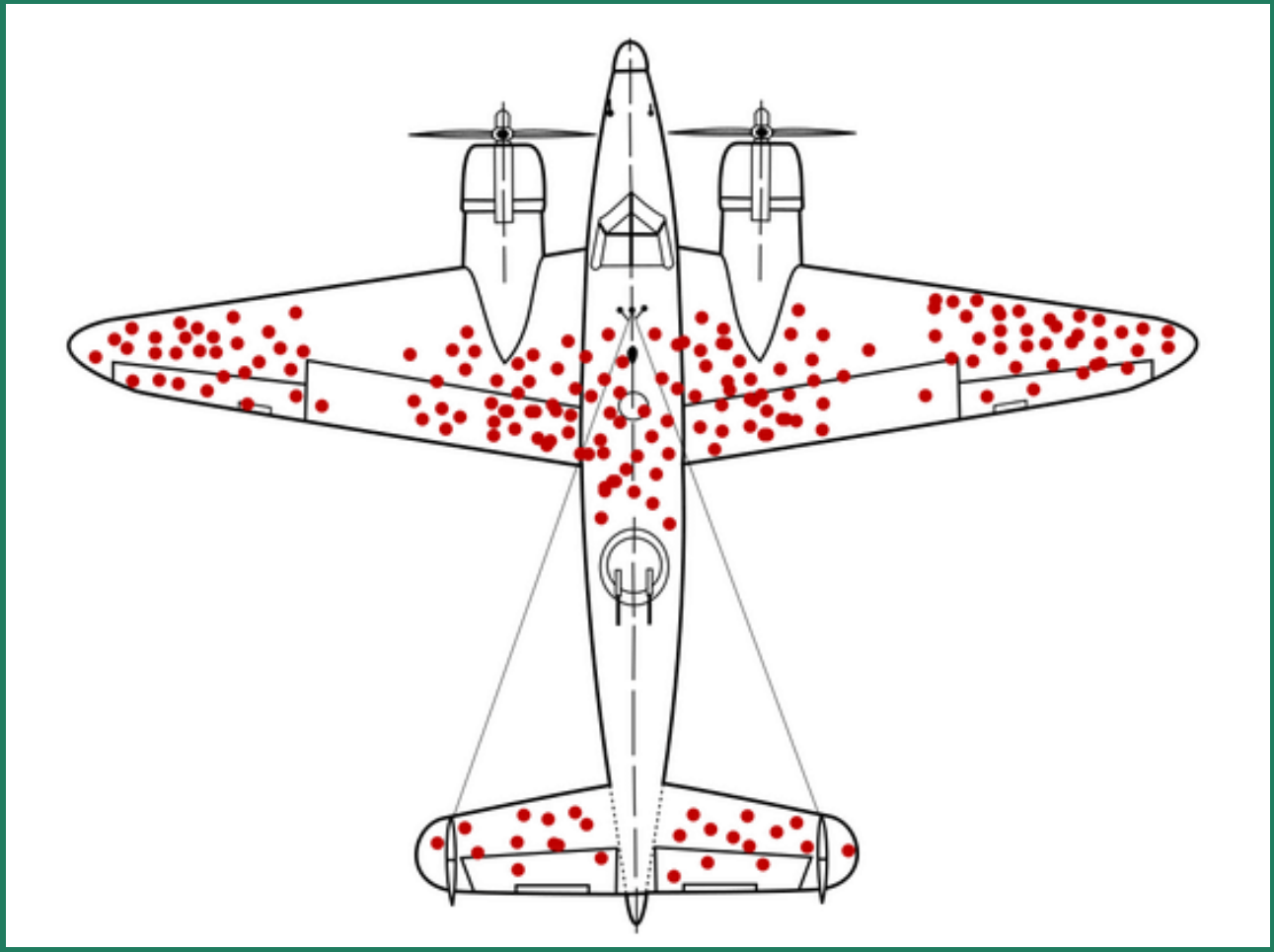
假设你正在设计第一个按揭计划。

如果将利率保持在3.85%，你的市场份额为53.1%
如果将利率提高到4.85%，你的市场份额下降到25.7%
如果将利率提高到5.85%，你的市场份额下降到9.5%
如果将利率降低到2.85%，你的市场份额增加到78.7%

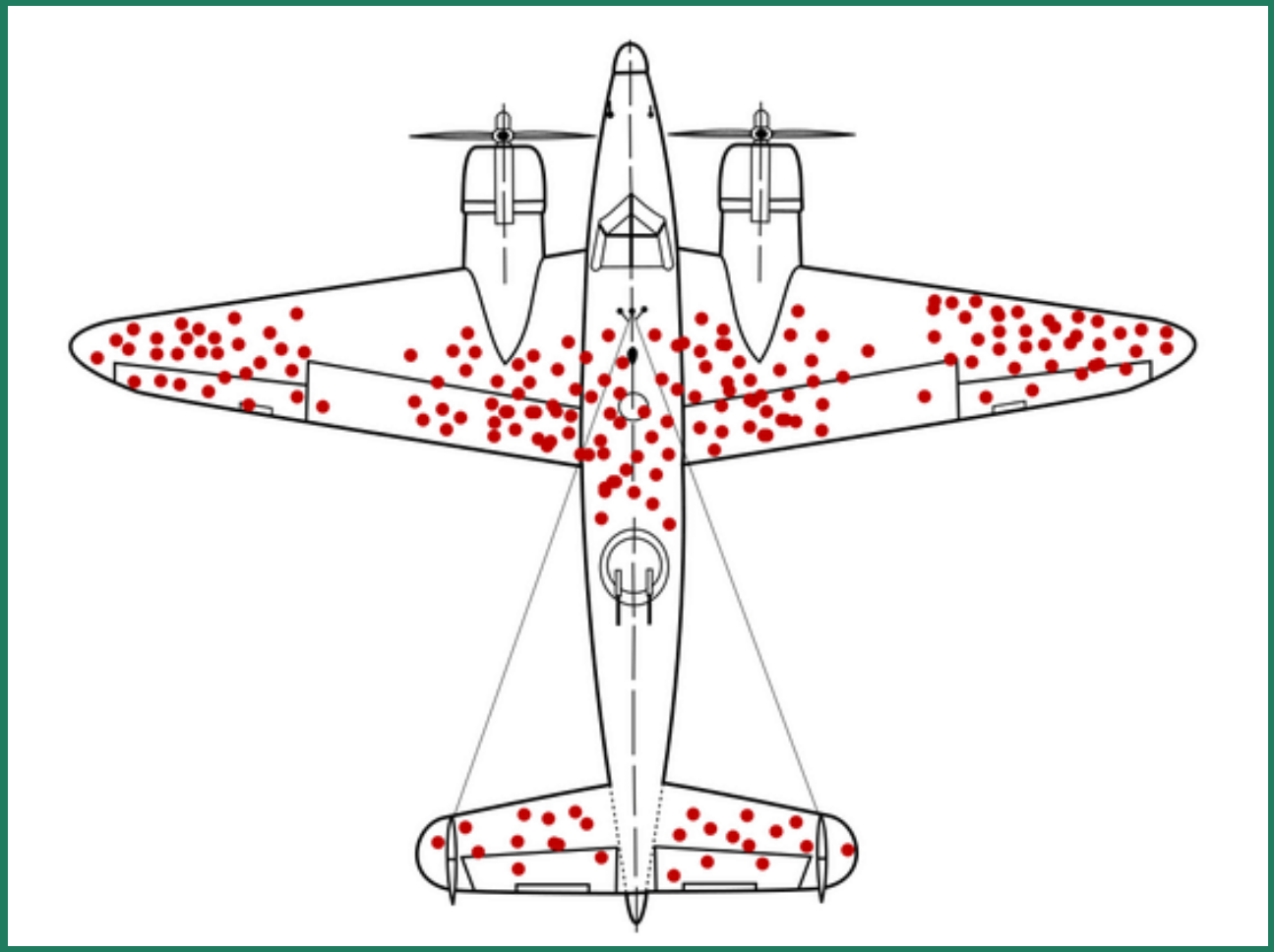
你可以选择一个平衡利润和市场份额的利率！

Selection Bias

选择偏差



In WWII, some planes never come back, and some come back with bullet holes. Here is the distribution of bullet holes. How would you reinforce the plane to increase the survival rate?



在第二次世界大战中，一些飞机坠毁了，而一些则带着弹孔返回。以下是弹孔的分布。你会如何加强飞机以提高生存率？

You should reinforce places where there are no holes, because if the plane will never come back when it has holes in these places! This is called the survivalship bias.

We only see part of the dataset. The other part is missing!

When you survey existing consumers, you don’t know why consumers are not buying your products. You don’t get honest opinion on your product development.

你应该加强没有弹孔的地方，因为如果这些地方有弹孔，飞机就不会回来！这被称为幸存者偏差。

我们只看到了数据集的一部分。另一部分是缺失的！

当你调查现有消费者时，你不知道为什么消费者不购买你的产品。你无法获得对产品开发的真实反馈。

Credit risk modeling

Based on whether a consumer repays the loan, you build an algorithm that predicts consumer repayment.

However, you only get data on consumers who got your loan, but do not get data on consumers who did not get your loan.

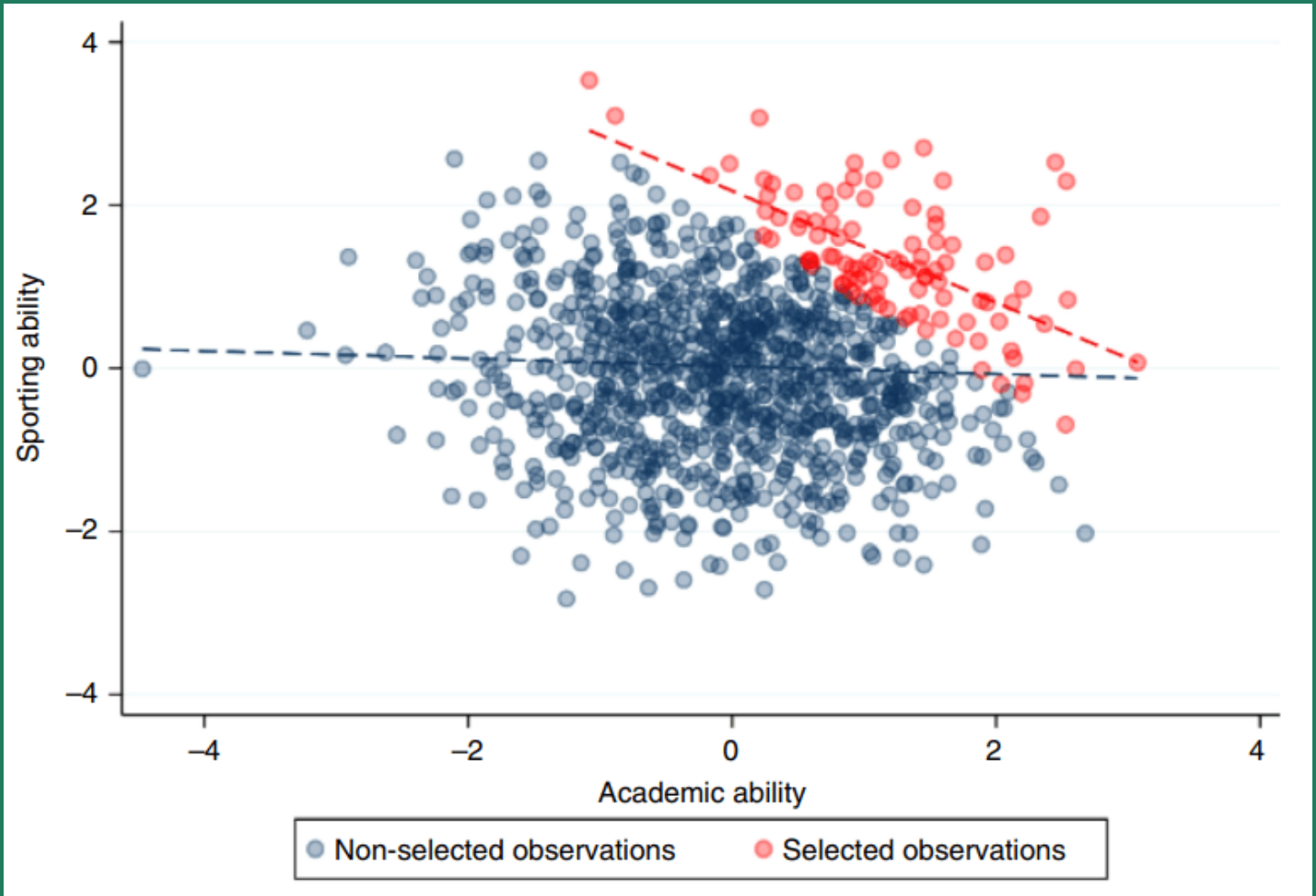
Some financial institutions randomly extend loans to some consumers who they predict will not repay.

信用风险建模

基于消费者是否偿还贷款，你建立一个算法来预测消费者的还款能力。

然而，你只能获得获得贷款的消费者的数据，却无法获得未获得贷款的消费者的数据。

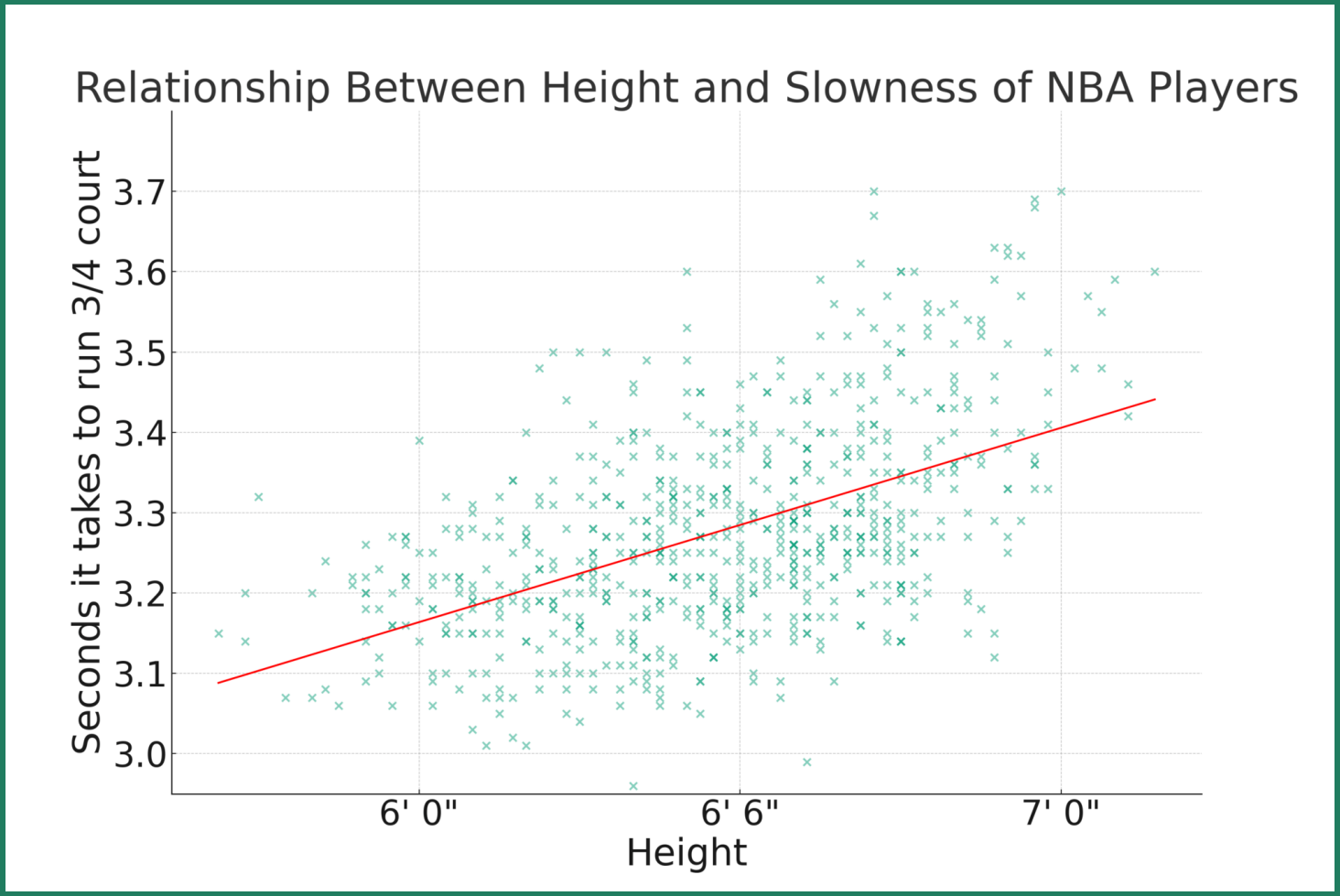
一些金融机构会随机向一些他们预测不会还款的消费者发放贷款。

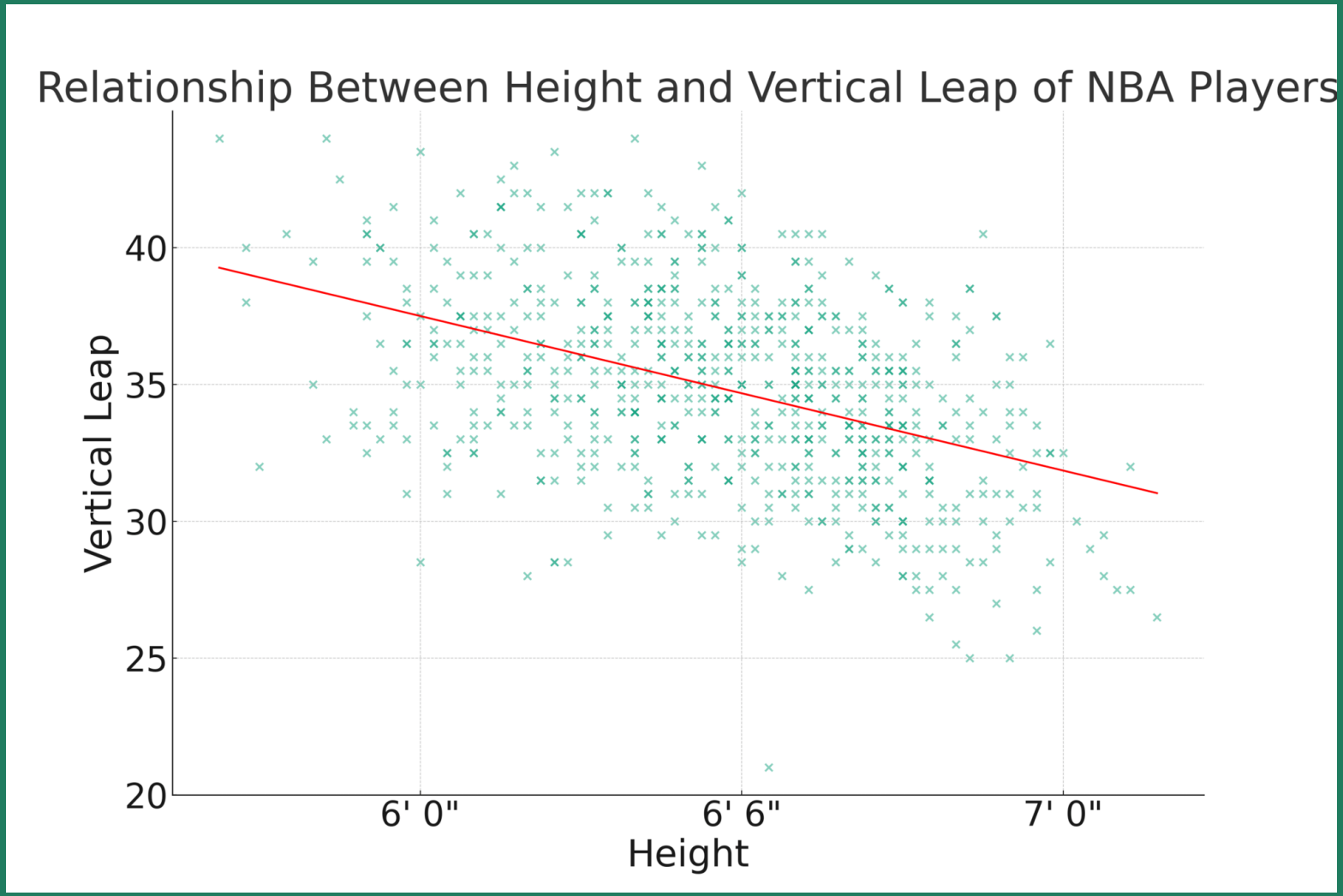


运动能力与学术能力是否存在负相关关系?

店铺位置是否与管理技能负相关（你关闭经营不善的店铺）?

再以篮球为例，如果只看 NBA 球员，会发现身高比较高的人得分率反而不如身高矮的人得分率高。这是因为身高矮还能进 NBA 的人必然是用其他优势补足了身高的弱势。




Relationship Between Height and Vertical Leap of NBA Players
Vertical Leap
40
35
30
25
20
6' 0"
6' 6"
7' 0"
Height

课后讨论问题：

1. 举一个你身边的选择偏差例子
2. 我们讲的定价中，所有消费者的价格都是相同的。但现实中，我们经常会有“一人一价”，根据数据对不同消费者收取不同的价格。如果你用大数据为你的产品制定个性化定价，你会收集哪些数据？你会如何定价？