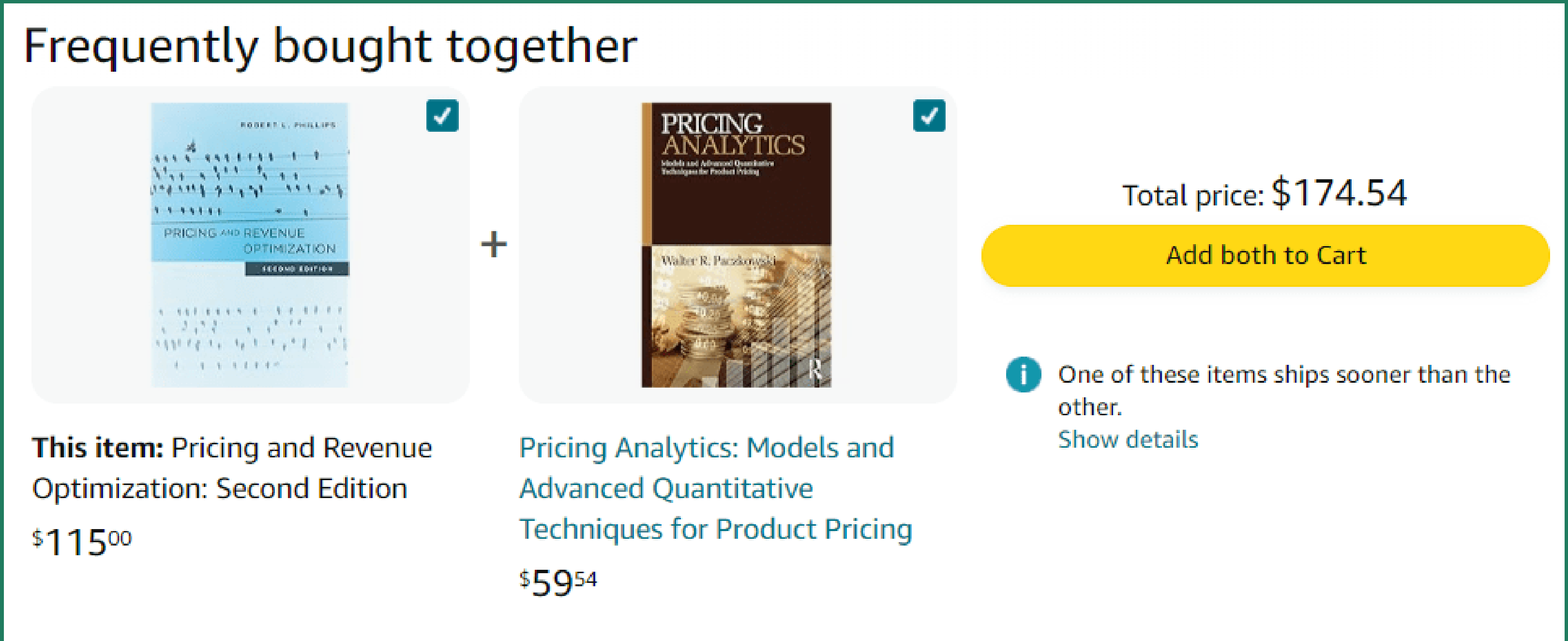
Market Basket Analysis

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

ROBERT L. PHILLIPS
PRICING AND REVENUE OPTIMIZATION
SECOND EDITION
☑
+
PRICING ANALYTICS
Models and Advanced Quantitative Techniques for Product Pricing
Walter R. Paczkowski
☑

This item: Pricing and Revenue Optimization: Second Edition

$115.00

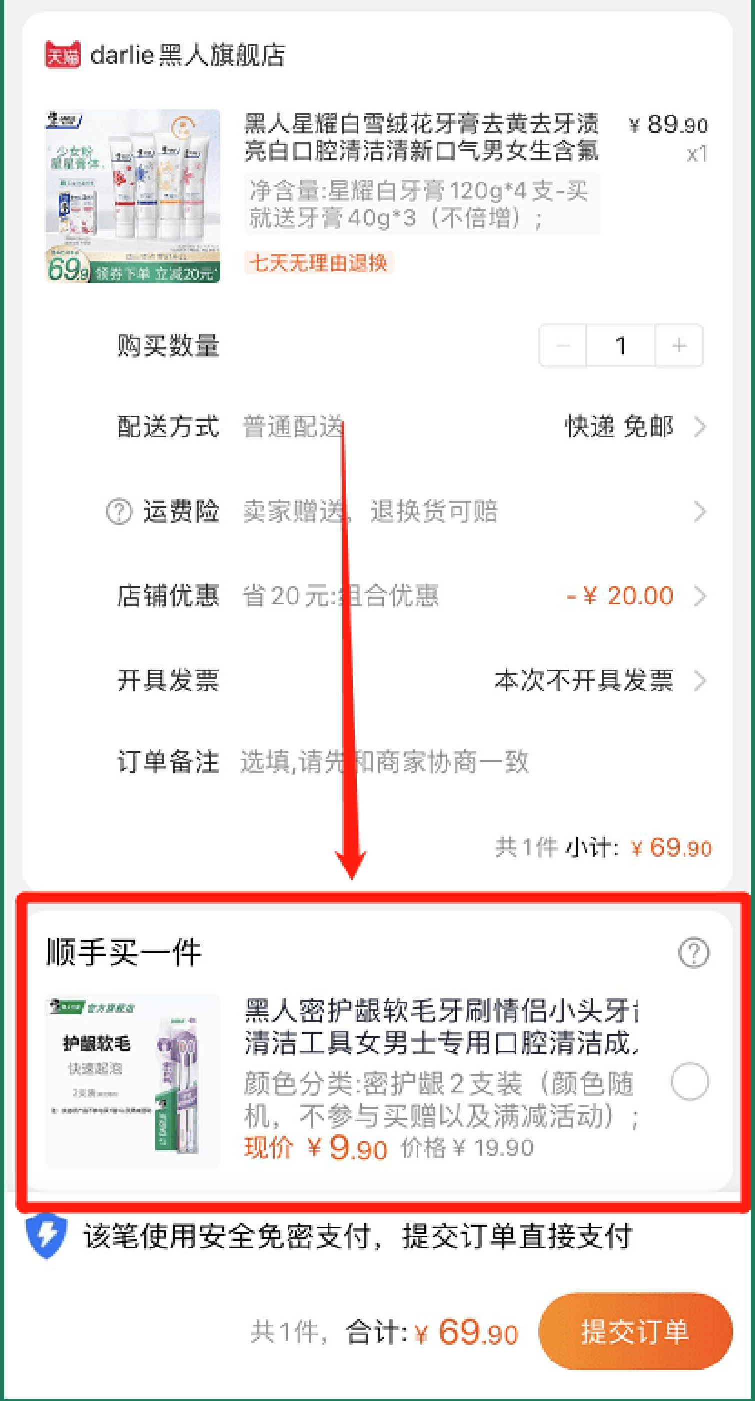
Pricing Analytics: Models and Advanced Quantitative Techniques for Product Pricing

$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

黑人星耀白雪绒花牌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刷清洁工具女男士专用口腔清洁成人

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

An association is the co-occurrence of two or more things. Beers may be associated with diaper, potato chips, or nuts.

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

关联是两个或更多事物的共同出现。啤酒可能与尿布、薯片或坚果相关联。

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的条件。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 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

Confidence 被定义为当左边产品被购买的情况下，右边产品被购买的概率.

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

换句话说, 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($X \rightarrow Y$) is not always equal to confidence($Y \rightarrow X$), for instance:

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

confidence(Bachelor → 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， $\text{confidence}(X \to Y)$ 并不总是等同于 $\text{confidence}(Y \to X)$, 举例来说：

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

$\text{confidence}(\text{Bachelor} \to \text{MBA}) = 0.05$: 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

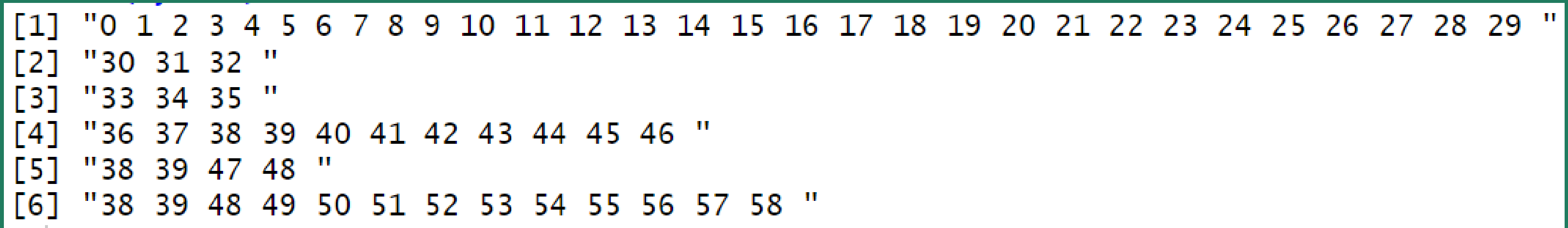
一个更重要的指标, lift, 表示两组产品同时购买和分别购买之间的关系:

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

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.

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

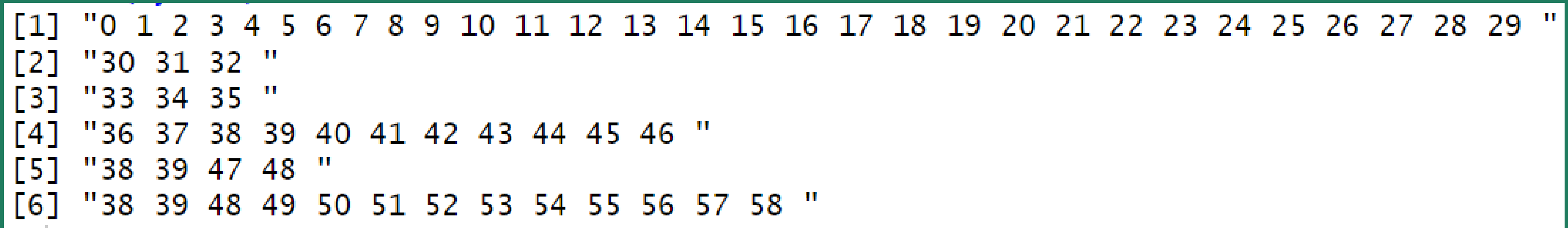


```
[1] "0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 "
[2] "30 31 32 "
[3] "33 34 35 "
[4] "36 37 38 39 40 41 42 43 44 45 46 "
[5] "38 39 47 48 "
[6] "38 39 48 49 50 51 52 53 54 55 56 57 58 "
```

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

我们首先载入数据

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

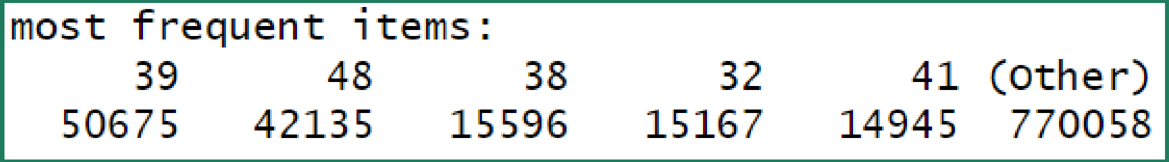


```
[1] "0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 "
[2] "30 31 32 "
[3] "33 34 35 "
[4] "36 37 38 39 40 41 42 43 44 45 46 "
[5] "38 39 47 48 "
[6] "38 39 48 49 50 51 52 53 54 55 56 57 58 "
```

第二个消费者一次同时购买了编号为 30, 31, 和 32 的产品

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

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

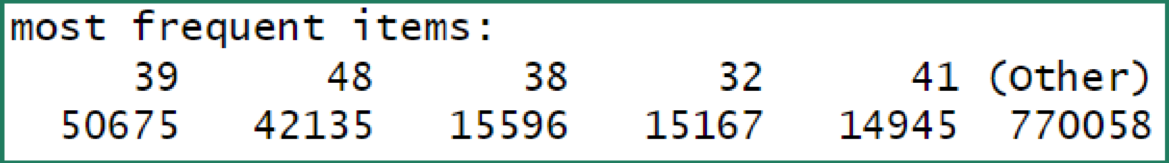


```
most frequent items:
     39      48      38      32      41 (Other) 
  50675   42135   15596   15167   14945  770058 
```

These are the most popular items in the transaction records.

```r
mydata = strsplit(mydata, " ")
transactions <- as(mydata, "transactions")
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 
```

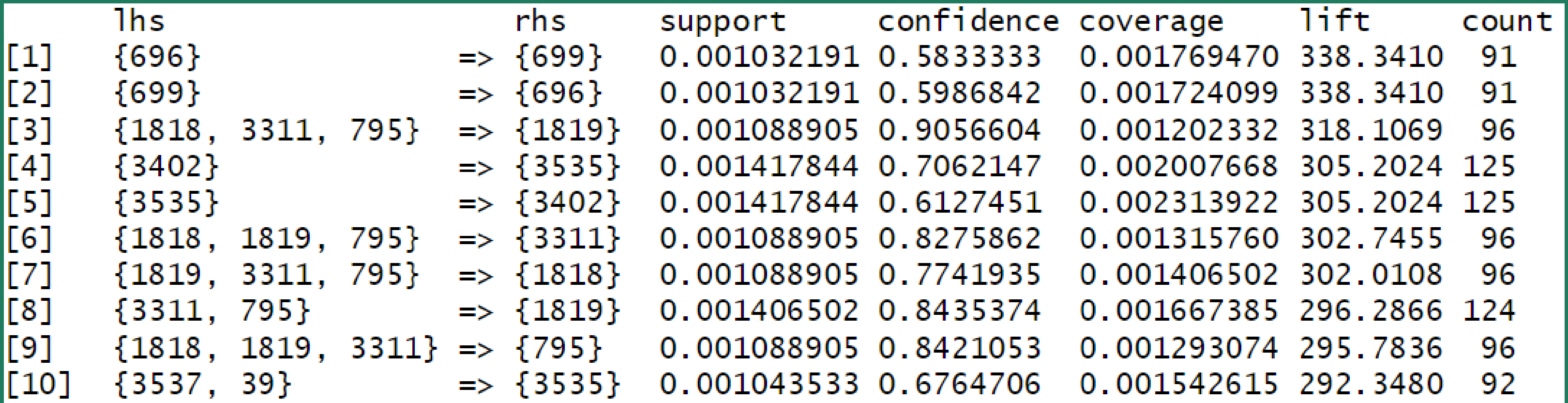
这些是交易记录中最受欢迎的物品。

```r
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 $\{A\} \to \{B\}$, 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.

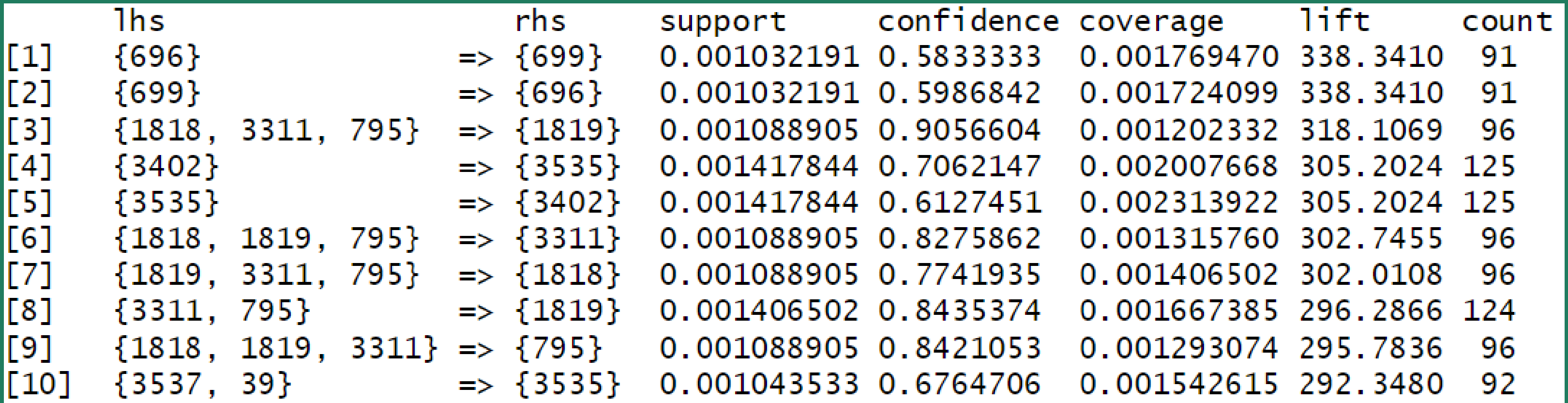
```
rules <- apriori(transactions,
                 parameter= list(supp=0.001, conf=0.4))
inspect(sort(rules, by="lift"))
```

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) support 至少为0.001，confidence 至少为0.4。然后，我们根据 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?"

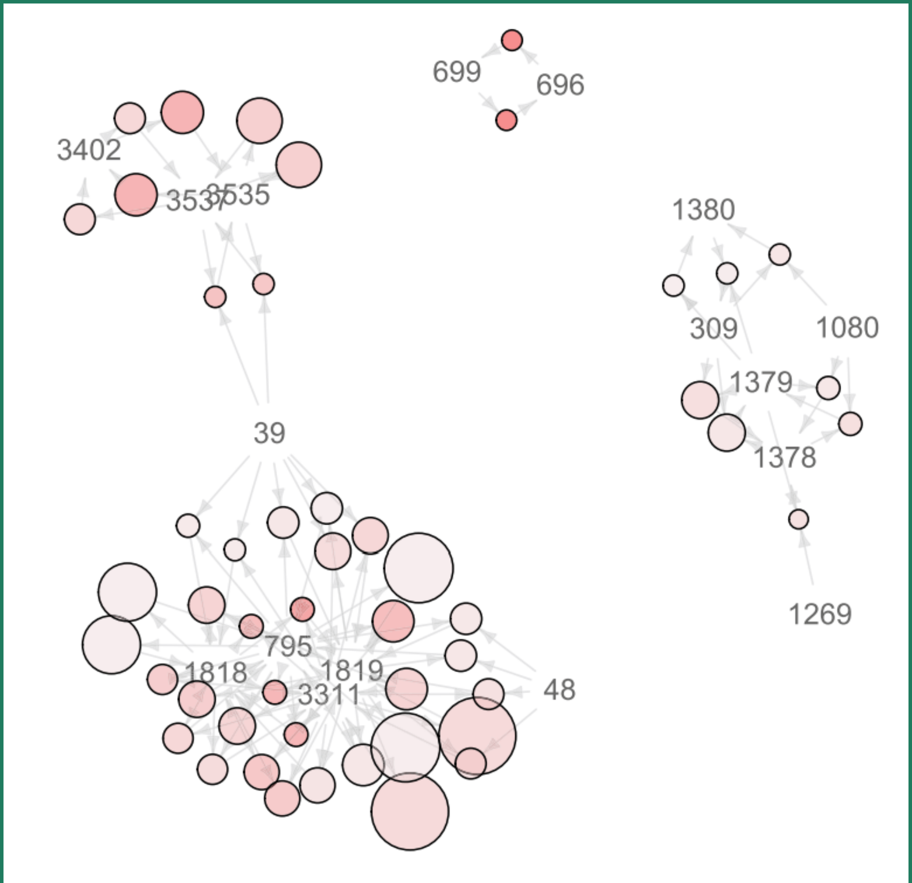


```
     lhs                    rhs    support     confidence coverage    lift     count
[1]  {696}               => {699}  0.001032191 0.5833333  0.001769470 338.3410  91
[2]  {699}               => {696}  0.001032191 0.5986842  0.001724099 338.3410  91
[3]  {1818, 3311, 795}   => {1819} 0.001088905 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？”

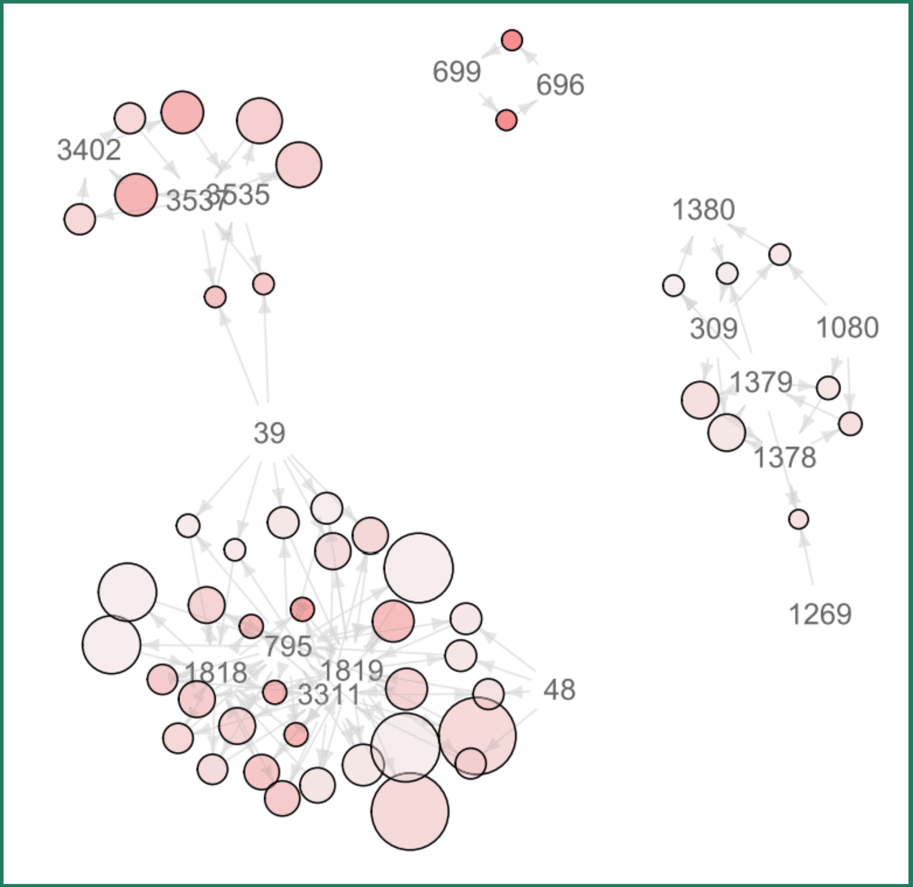
```r
plot(rules , method="graph", control= list(type="items"))
```

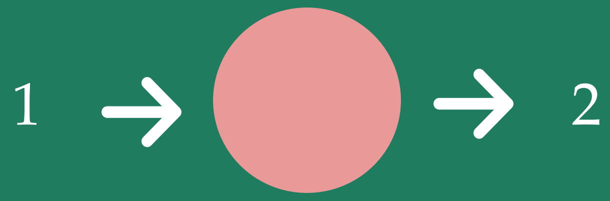
We can further visualize the rules we have detected. You will get something like this (it varies with different for the system):



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

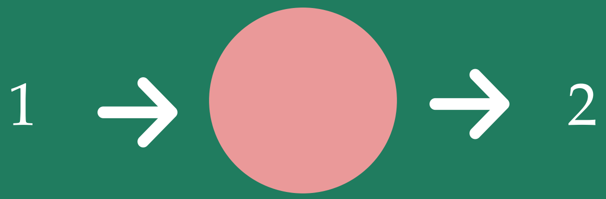
我们可以进一步将我们检测到的规则进行可视化展示。您将会看到类似以下内容的展示（实际展示会因系统不同而有所变化）：





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}。

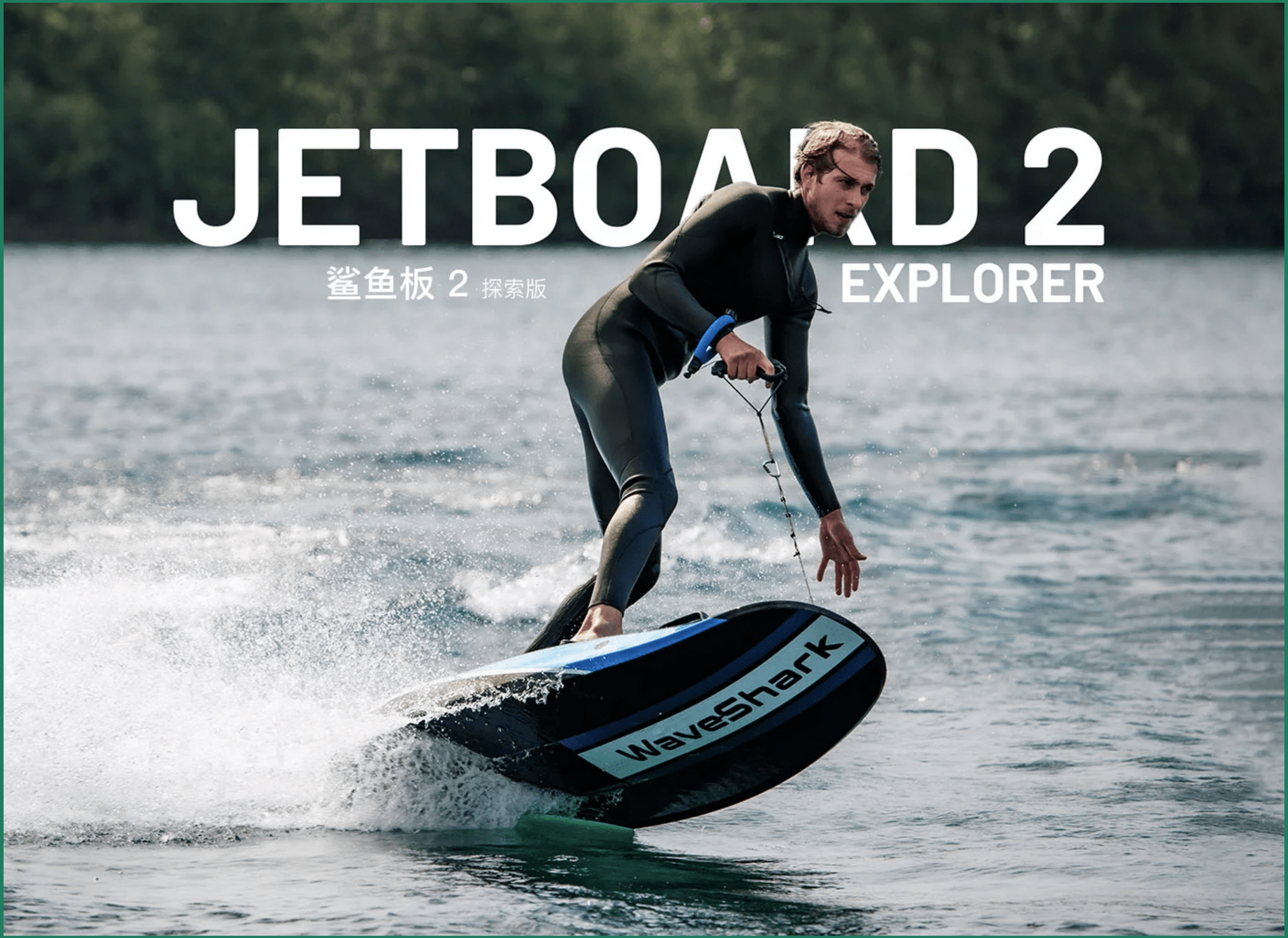
圆的大小（面积）代表规则的 support，颜色深浅代表 lift（颜色越深表示 lift 越高）。

The complete code is here:

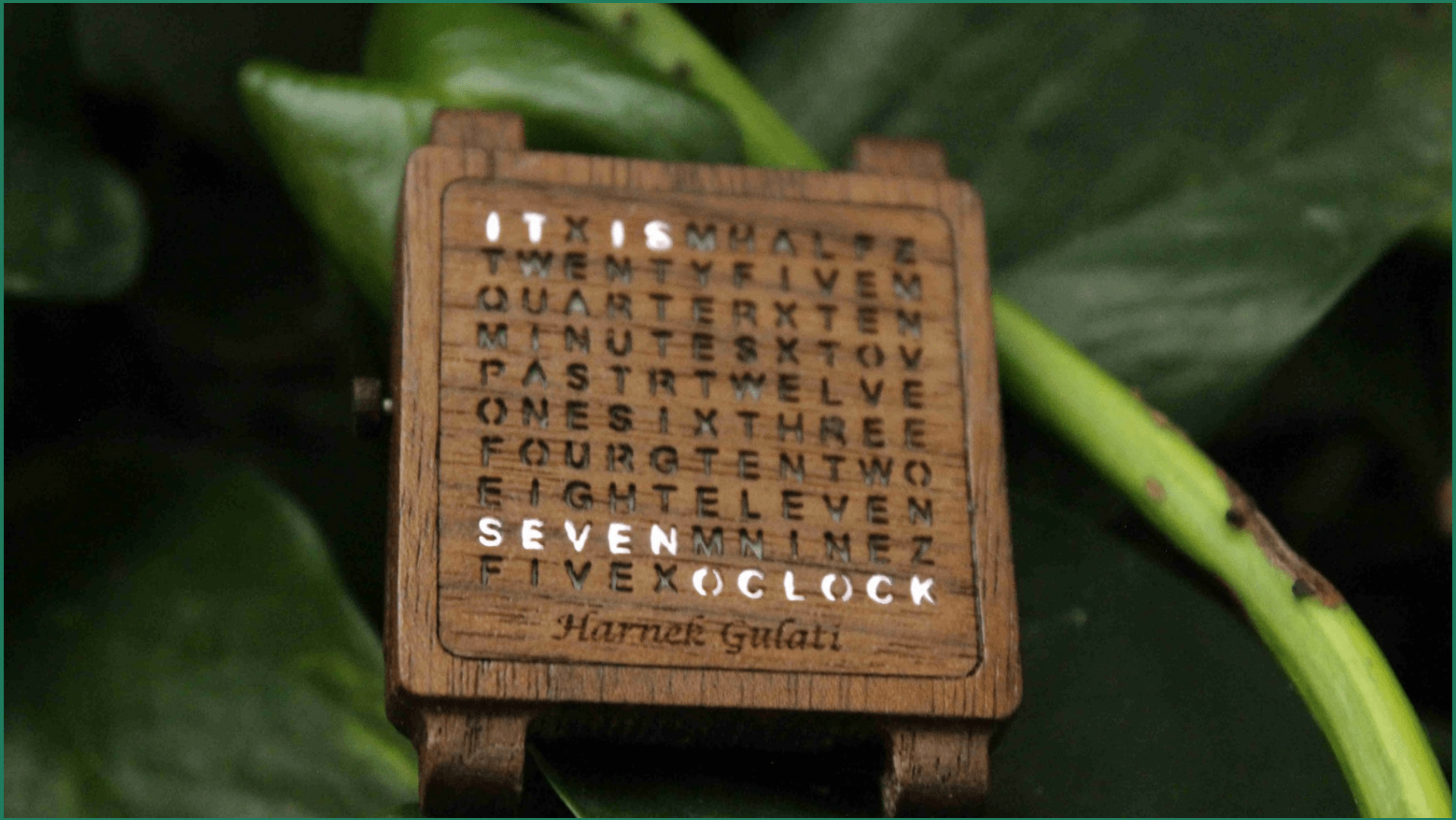
```r
library(arules)
library(arulesViz)
mydata = readLines("https://ximarketing.github.io/data/basket.txt")
mydata = strsplit(mydata, " ")
transactions <- as(mydata, "transactions")
rules <- apriori(transactions,
    parameter= list(supp=0.001, conf=0.4))
inspect(sort(rules, by="lift"))
plot(rules ,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

Guess the Price

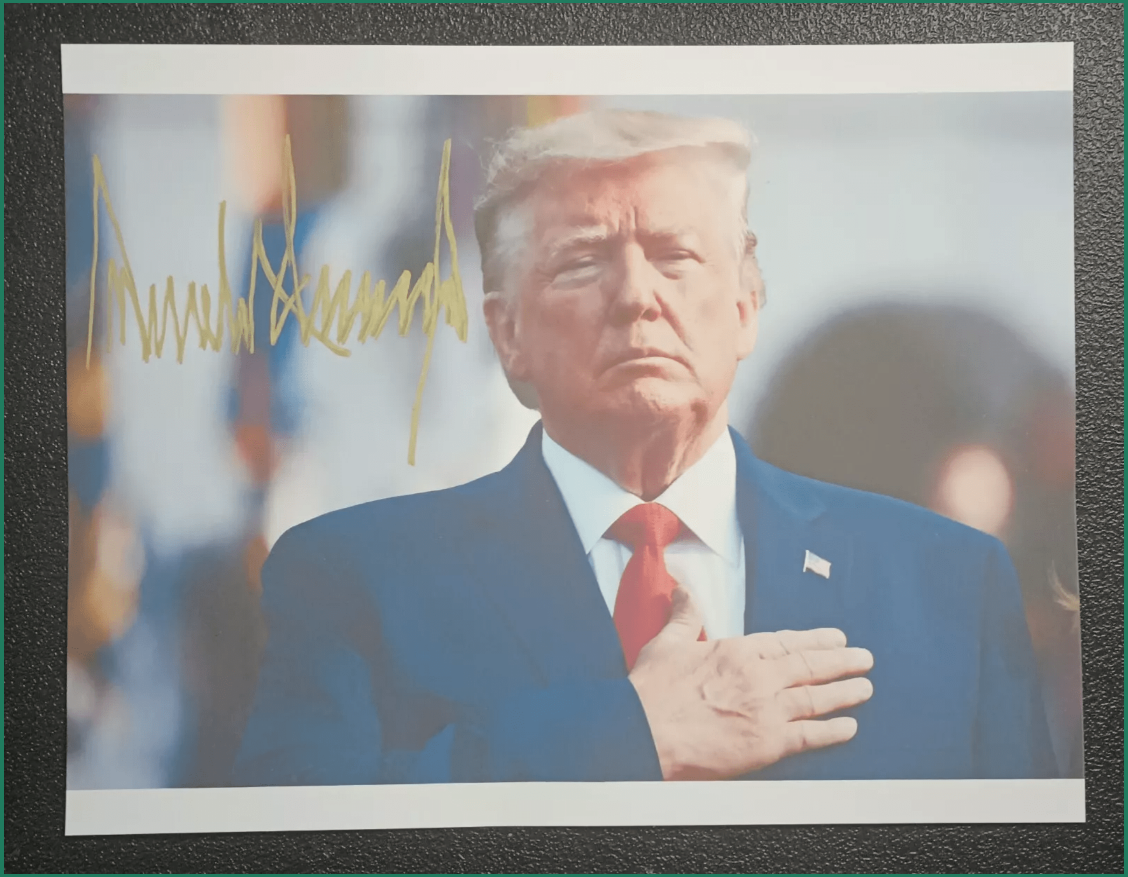
猜一猜价格



电动冲浪板 (electric surfboards)



木制手表 (Wooden Watch)



特朗普签名 (Donald Trump Autograph)

Price Sensitivity Analysis

价格敏感性分析

Question 问题

Suppose that you are developing a completely new-to-the-world product that no consumers are familiar with. How would you determine the price for the product?

假设你正在开发一款全新的世界首创产品，没有消费者熟悉它。你会如何确定该产品的价格?

Can we conduct a survey?

One simple approach is to just ask consumers for prices. However, this would result in many unrealistic prices. For instance, in the past, a firm asked consumers about the prices for a shampoo, and many consumers stated prices $0 and $75, which are totally unreasonable.

For a new product, consumers do not have an anchor and need some guide in determining the price.

我们能进行一项调查吗?

一个简单的方法就是询问消费者价格。然而，这样会导致很多不现实的价格。例如，过去有家公司向消费者询问洗发水的价格，许多消费者给出了0美元和75美元的价格，这是完全不合理的。

对于新产品，消费者没有参照点，需要一些指导来确定价格。

van Westerndorp method

Instead of deciding on the optimal price point, the method looks for the optimal price range.

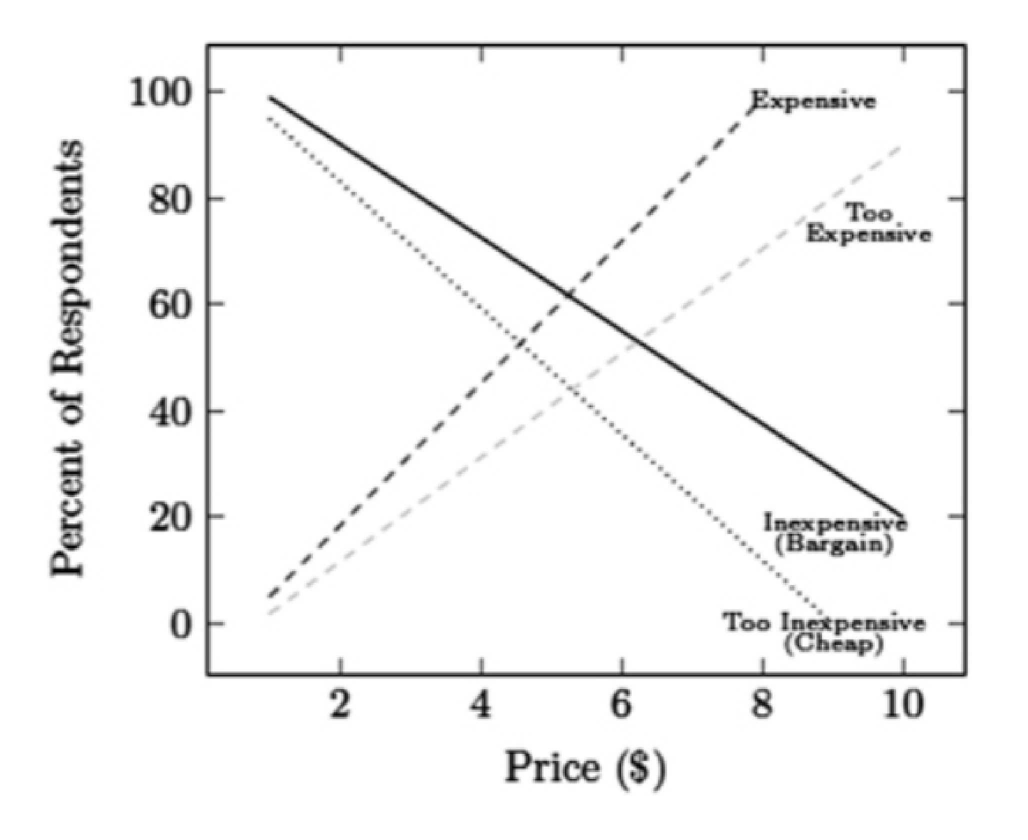
First, provide a list of feasible prices, e.g., from $50 to $500 with $50 increments. Then, ask respondents to select one of the price for the following four questions:

1. At what price the product is too expensive you would not buy it?
2. At what price is the product becoming expensive so you would have to think about buying it?
3. At what price is the product a bargain, great value for the money?
4. At what price is the product so inexpensive you would feel the quality is questionable?

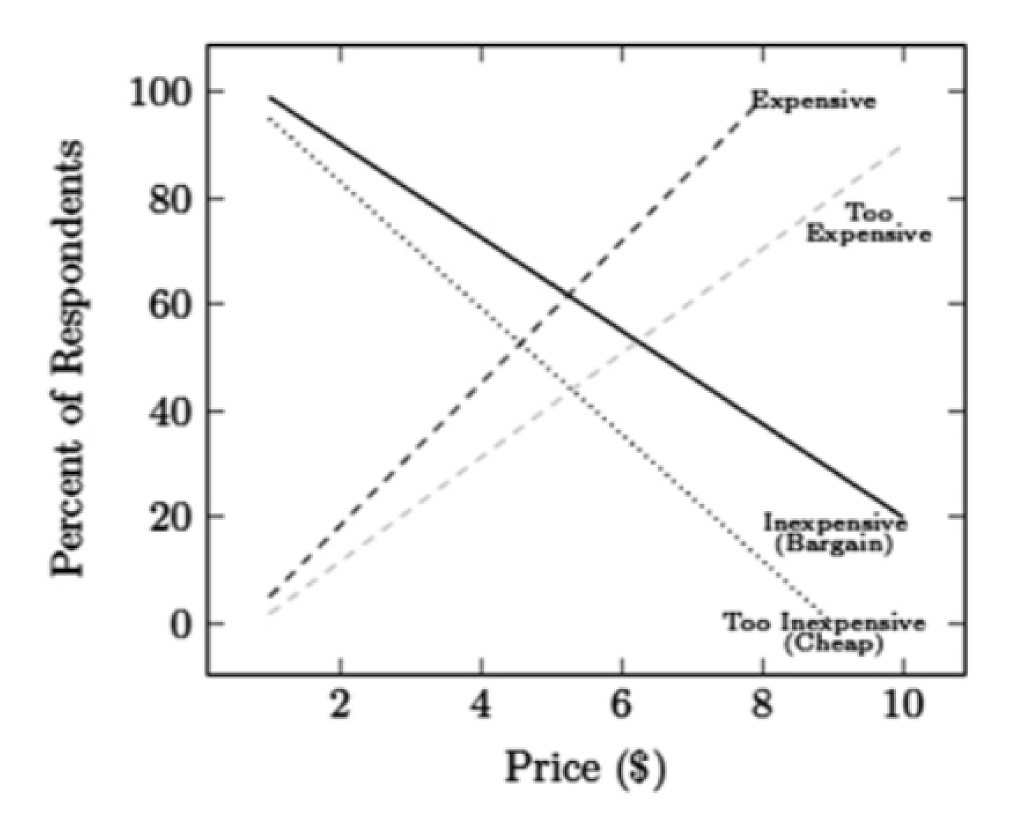
与决定最佳价格点相反，这种方法寻找最佳价格范围。

首先，提供一个可行价格列表，例如，从 50 元到 500 元，每 50 元递增。然后，要求受访者为以下四个问题选择一个价格。

1. 产品定价多少时，你会觉得太贵，不会购买?
2. 产品定价多少时，你会觉得有点贵，需要考虑是否购买?
3. 产品定价多少时，你会觉得是物超所值，性价比很高?
4. 产品定价多少时，你会觉得太便宜，对产品质量产生质疑?



We can then draw four lines. For instance, the expensive line illustrates "how many people view this price (or a lower price) as expensive"?



然后我们可以画四条线。例如，昂贵线 (expensive) 表示“多少人认为这个价格（或更低的价格）太贵”?

There are two key prices points:

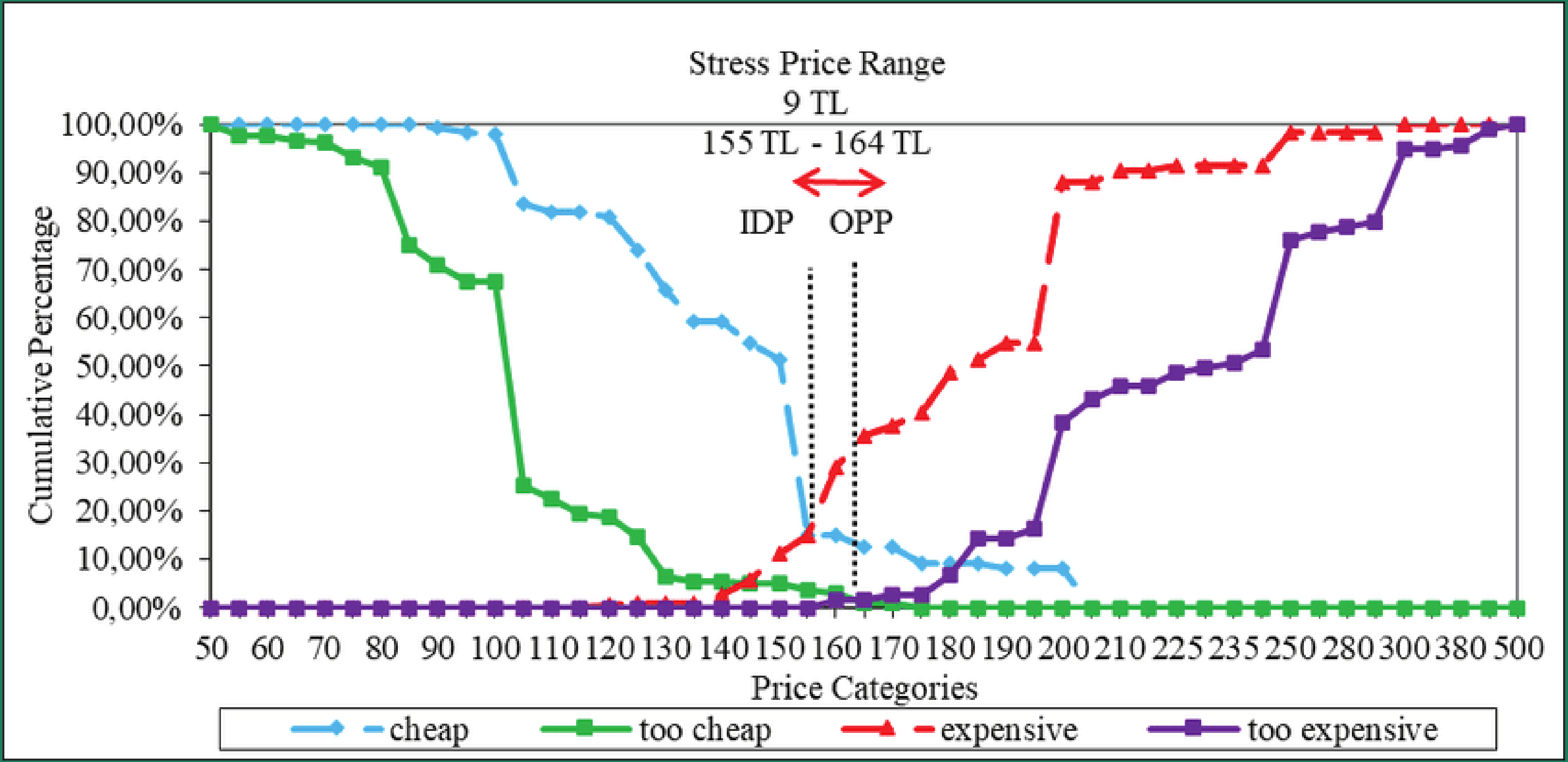
1. Optimal Price Point (OPP): The price at which the number of consumers who rated "too expensive" equals consumers rating it "cheap", i.e., Too Expensive = Cheap.
2. Indifference Price Point (IDP): The price at which the number of consumers who rated "expensive" equals consumers rating it "bargain" i.e., Expensive = Bargain.

The optimal price range lies between the above two prices.

这里有两个关键价格点：

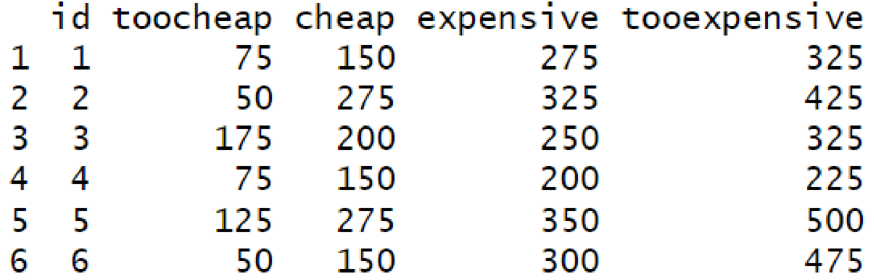
1. Optimal Price Point (OPP) 最佳价格点: 认为“太贵”与认为“便宜”的消费者数量相等，即，Too Expensive = Cheap
2. Indifference Price Point (IDP) 中立价格点: 认为“昂贵”与认为“便宜”的消费者数量相等, 即 Expensive = Bargain.

最佳价格范围位于上述两个价格之间。


Stress Price Range
9 TL
155 TL - 164 TL
IDP
OPP
Cumulative Percentage
100,00%
90,00%
80,00%
70,00%
60,00%
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 225 235 250 280 300 380 500
Price Categories
cheap
too cheap
expensive
too expensive

As before, we load data from the web:

```r
library(pricesensitivitymeter)
library(ggplot2)
mydata <- read.csv("https://ximarketing.github.io/data/VanWestendorp.csv")
head(mydata)
```

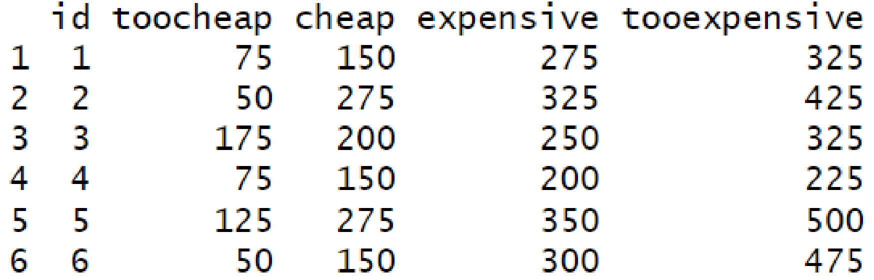


```
  id toocheap cheap expensive tooexpensive
1  1       75   150       275          325
2  2       50   275       325          425
3  3      175   200       250          325
4  4       75   150       200          225
5  5      125   275       350          500
6  6       50   150       300          475
```

Respondent 1 believes $75 is too cheap, $150 is a bargain, $275 is somewhat expensive whereas $325 is way too expensive.

与之前一样，我们从网络加载数据：

```r
library(pricesensitivitymeter)
library(ggplot2)
mydata <- read.csv("https://ximarketing.github.io/data/VanWestendorp.csv")
head(mydata)
```

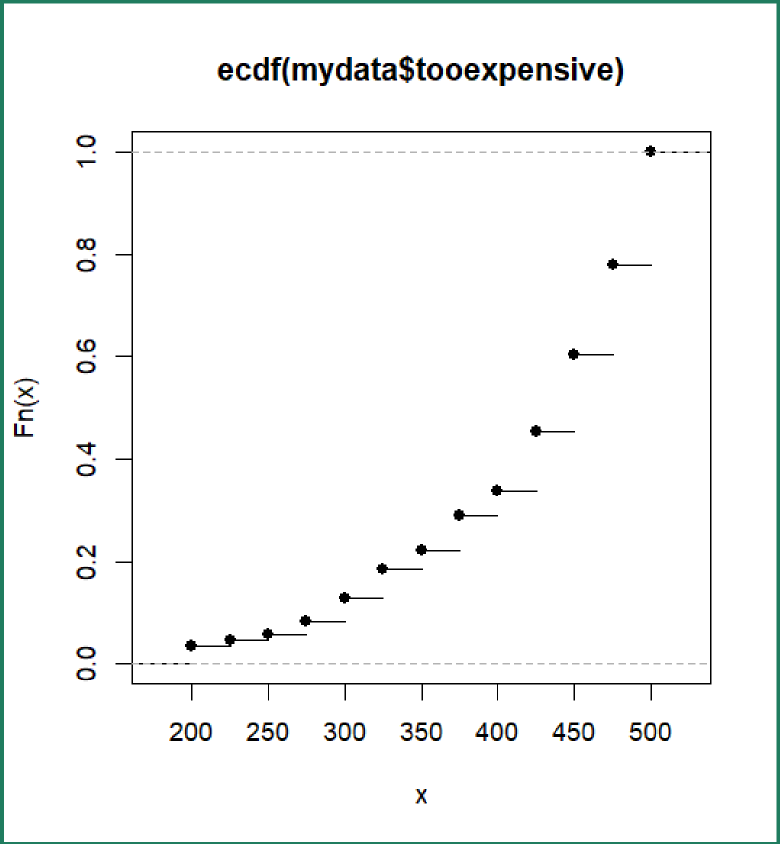


```
  id toocheap cheap expensive tooexpensive
1  1       75   150       275          325
2  2       50   275       325          425
3  3      175   200       250          325
4  4       75   150       200          225
5  5      125   275       350          500
6  6       50   150       300          475
```

受访者1认为75元太便宜，150元是个便宜货，275元有点贵，而325元太贵了。

We can visualize the distribution of consumer price perception. We take "too expensive" as an example.

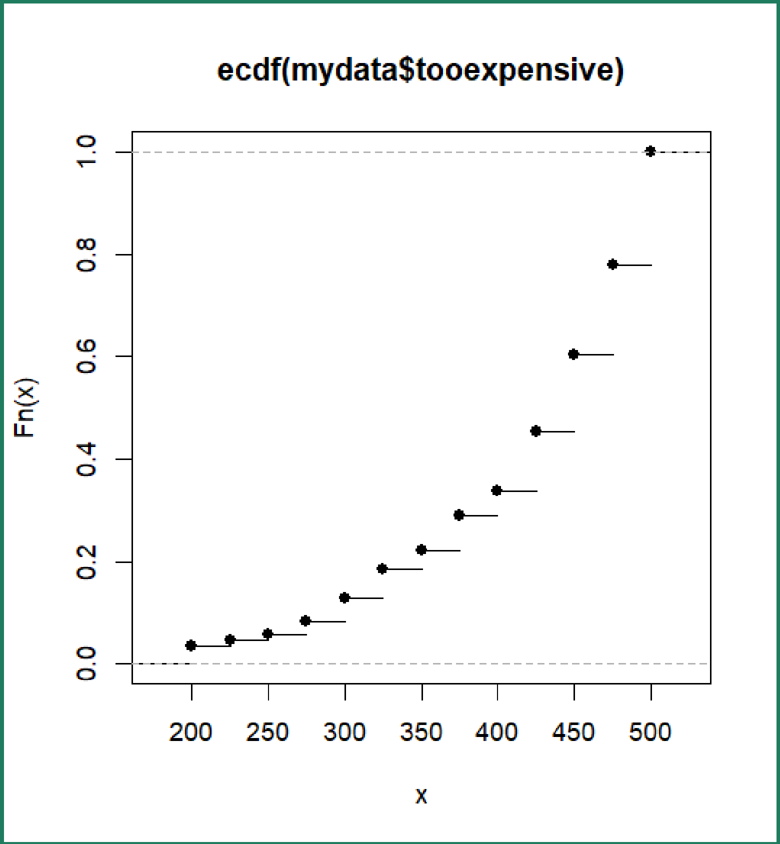
```r
plot(ecdf(mydata$tooexpensive))
```



Around 20% believe $325 is too expensive, and around 80% believe $475 is too expensive...

我们可以可视化消费者价格感知的分布。我们以“太贵”为例。

```r
plot(ecdf(mydata$tooexpensive))
```



大约20%的人认为325元太贵，大约80%的人认为475元太贵...

```
with(mydata, which(cheap <= toocheap | expensive <= cheap |
                     tooexpensive <= expensive))
mydata <- mydata[-51, ]
```

A reasonable answer should satisfy that too expensive > expensive > cheap > too cheap. Otherwise it does not make sense.

Here, we check which answer(s) does not satisfy the condition and remove it from our dataset.

```r
with(mydata, which(cheap <= toocheap | expensive <= cheap |
                     tooexpensive <= expensive))
mydata <- mydata[-51, ]
```

一个合理的答案应该满足太贵 > 贵 > 便宜 > 太便宜。否则就没有意义。 在这里，我们检查哪个答案不满足这个条件，并从数据集中删除它。

```r
psm <- psm_analysis(toocheap     = mydata$toocheap,
                    cheap        = mydata$cheap,
                    expensive    = mydata$expensive,
                    tooexpensive = mydata$tooexpensive,
                    validate = FALSE, interpolate = TRUE,
                    intersection_method = "median")
```

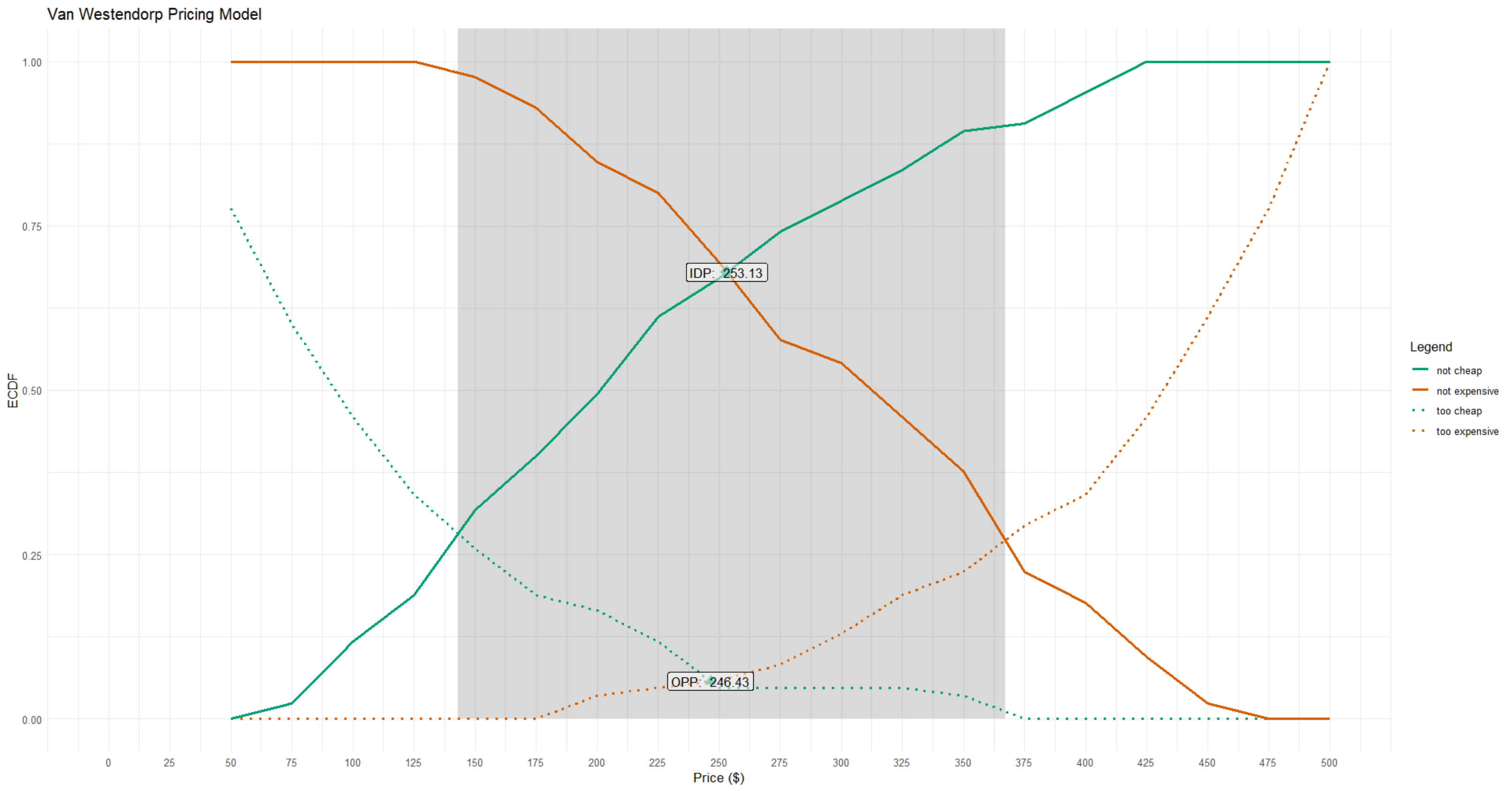
Next, we fit the data into our model. Here, PSM stands for "price sensitivity model."

接下来，我们将数据拟合到我们的模型中。在这里，PSM代表“价格敏感性模型”。

```r
scalebin <- 25
scalemax <- 500
psm_plot(psm) +
  scale_x_continuous(breaks=0:(scalemax/scalebin)*scalebin) +
  coord_cartesian(xlim=c(0, scalemax)) +
  theme_minimal() +
  ylab("ECDF") +
  xlab("Price ($)") +
  ggtitle("Van Westendorp Pricing Model")
```

Next, we specify the parameters for visualization, and plot the result using the tool provided by the package.

接下来，我们指定可视化参数，并使用该包提供的工具绘制结果。



Price range should be 246.43 to 253.13.

价格范围应该是246.43到253.13

The complete code is here:

```r
library(pricesensitivitymeter)
library(ggplot2)
mydata <- read.csv("https://ximarketing.github.io/data/VanWestendorp.csv")
psm <- psm_analysis(toocheap     = mydata$toocheap,
                    cheap        = mydata$cheap,
                    expensive    = mydata$expensive,
                    tooexpensive = mydata$tooexpensive,
                    validate = FALSE, interpolate = TRUE,
                    intersection_method = "median")
scalebin <- 25
scalemax <- 500
psm_plot(psm) +
  scale_x_continuous(breaks=0:(scalemax/scalebin)*scalebin) +
  coord_cartesian(xlim=c(0, scalemax)) +
  theme_minimal() +
  ylab("ECDF") +
  xlab("Price ($)") +
  ggtitle("Van Westendorp Pricing Model")
```

Conjoint Analysis

联合分析

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另一个有用工具，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

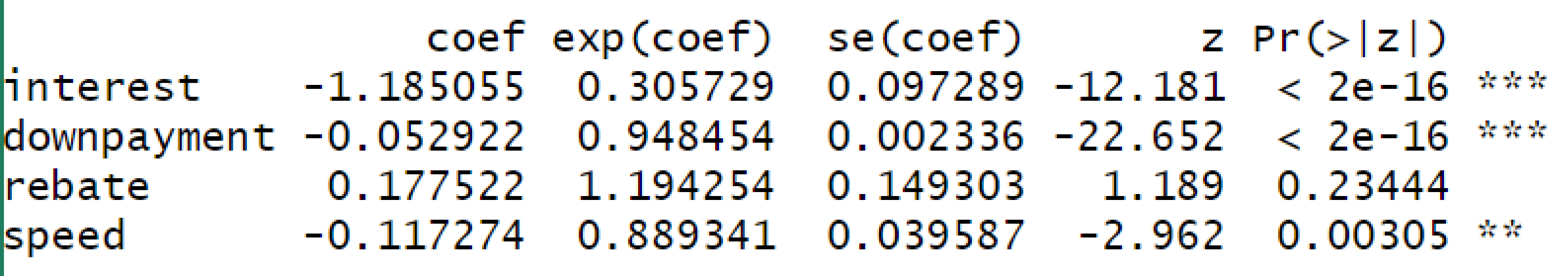
然后，您创建不同的组合，让消费者从类似以下的备选方案中进行选择：

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

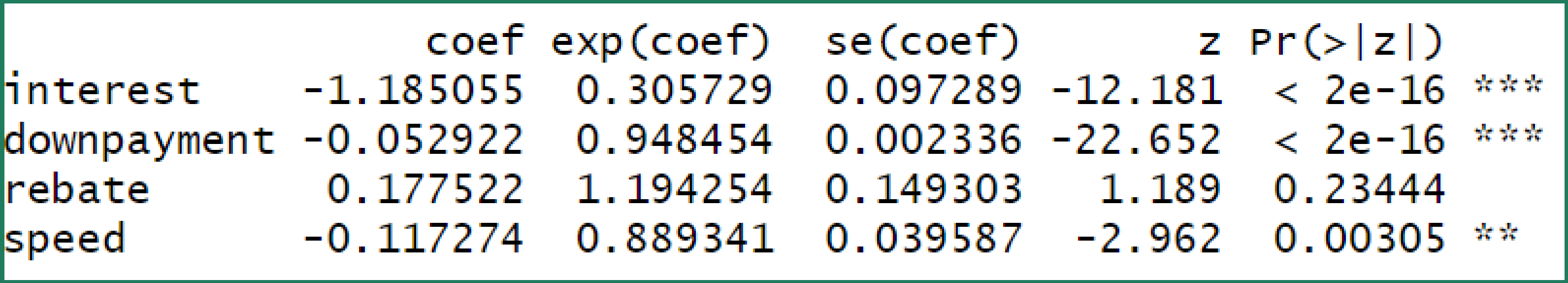
For example, you survey 6,000 consumers, and each consumer chooses among 3 alternatives. Then you plug the data into your conditional logistic model, and get results like this:



```
                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 Logistic 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!

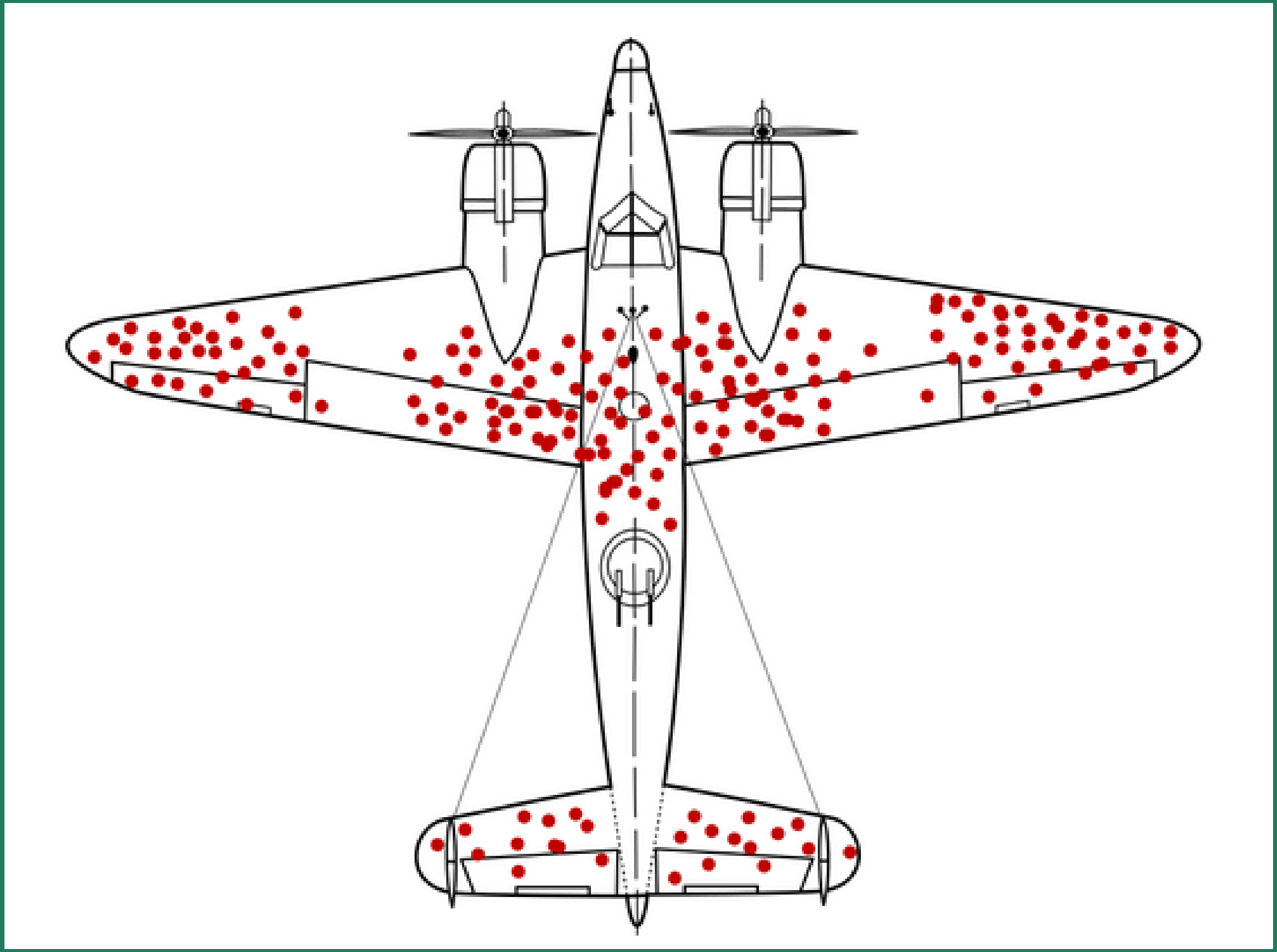
假设你正在设计第一个按揭计划。

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%

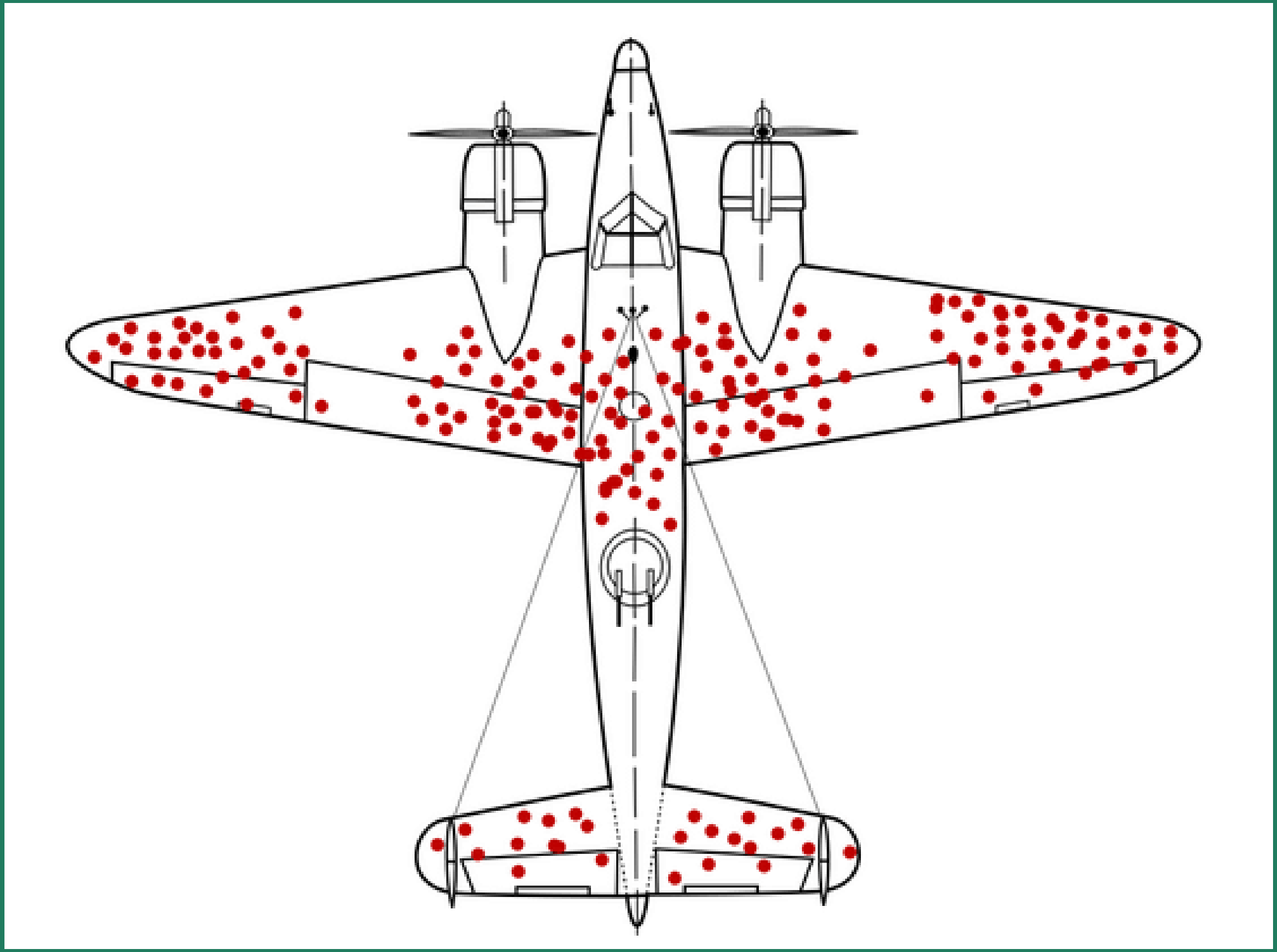
你可以选择一个平衡利润和市场份额的利率！

Selection Bias

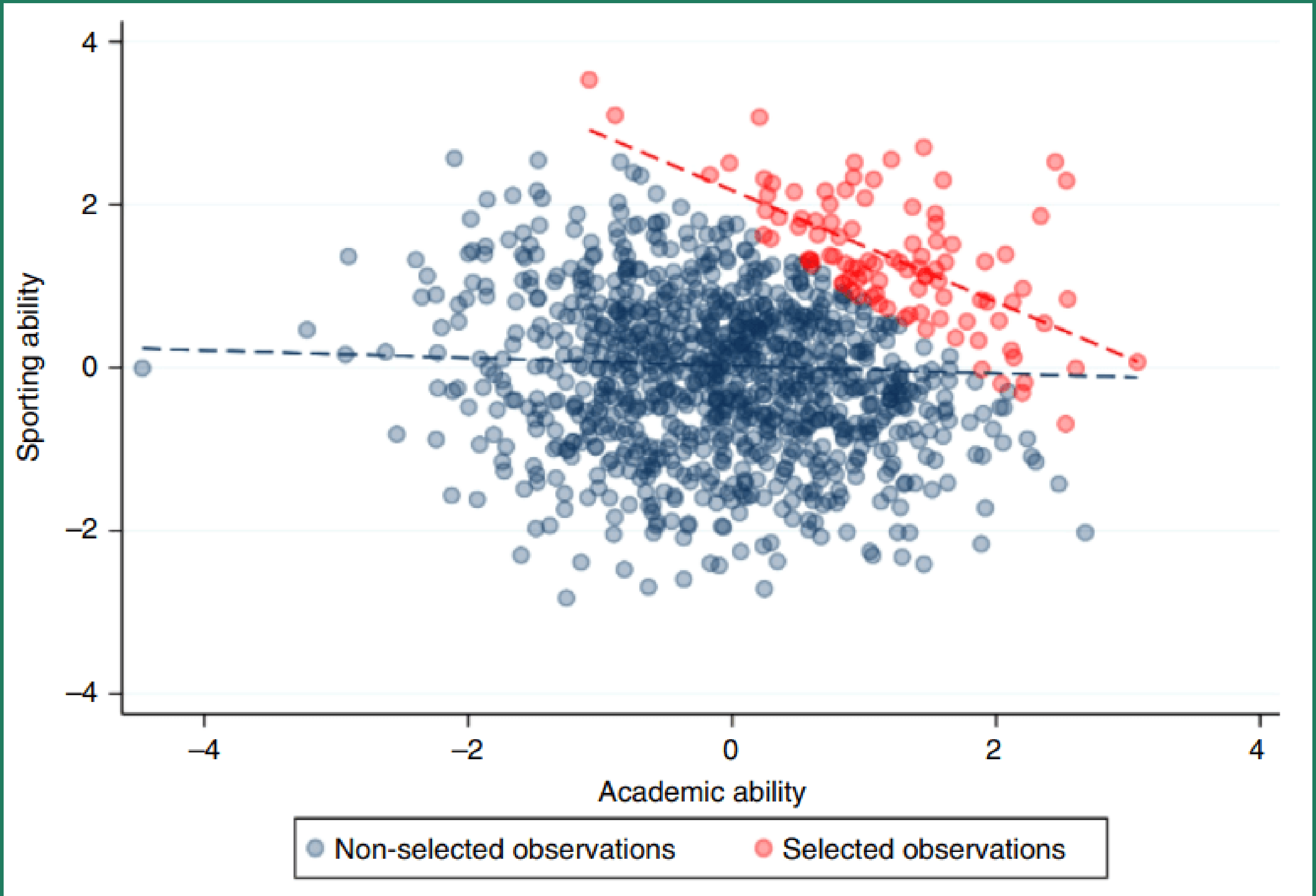
选择偏差



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?

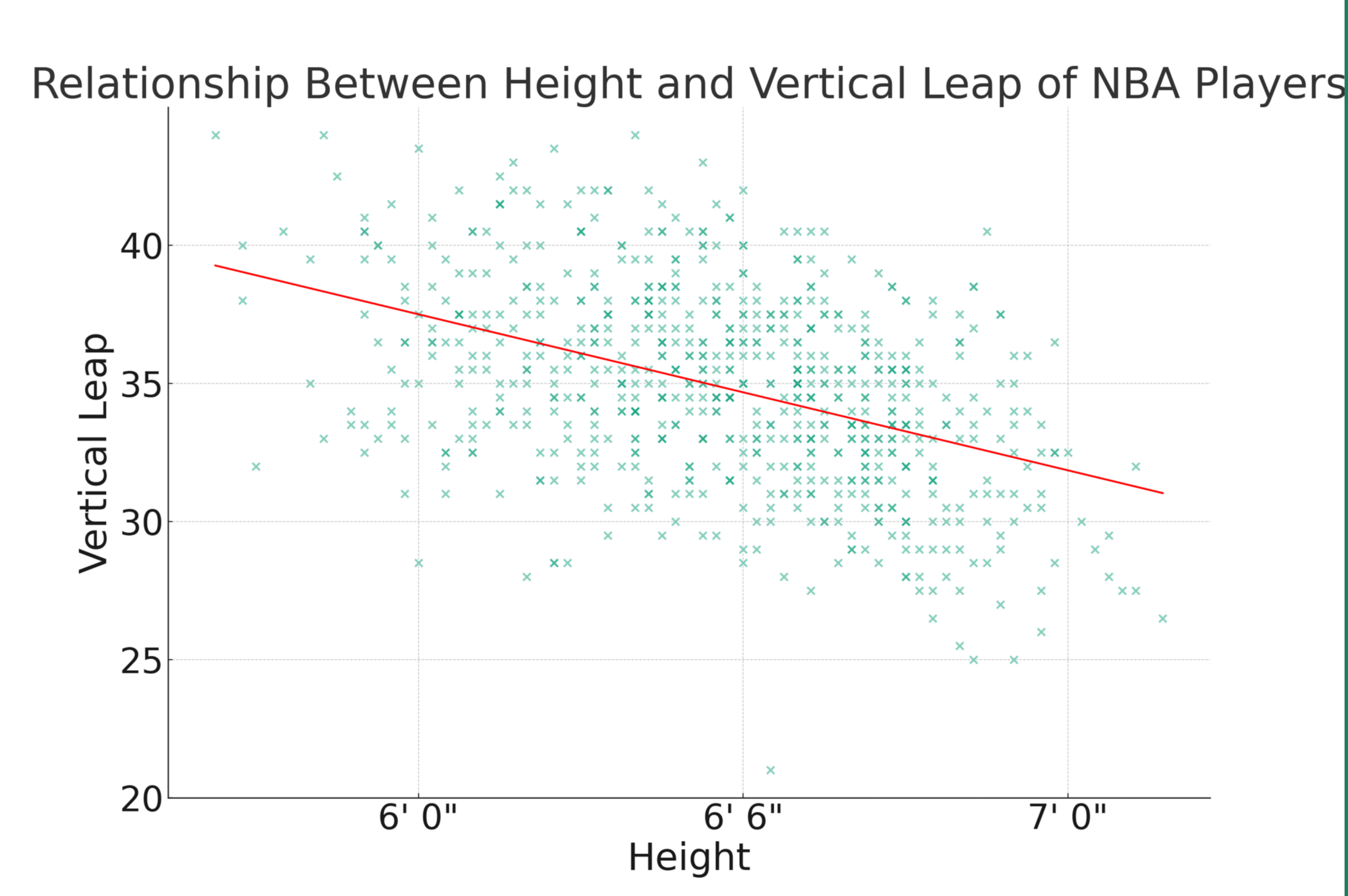


在第二次世界大战中，一些飞机坠毁了，而一些则带着弹孔返回。以下是弹孔的分布。你会如何加强飞机以提高生存率?



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）？


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