数据分析与决策技术丛书

HZ BOOKS
华章IT

Data Analysis Practice Based on EXCEL and Series of SPSS Tools

数据分析实战

基于EXCEL和SPSS系列工具的实践

纪贺元◎著

数据分析专家撰写，多年企业数据分析培训和咨询的经验结晶

实战性强，从“非科班出身”的企业数据分析人员的角度对商业数据分析进行了总结和归纳，运用大量事实和案例来展现“以业务为核心和抓手的数据分析”商业实践

机械工业出版社
China Machine Press

数据分析与决策
技术丛书

Data Analysis Practice Based on EXCEL and Series of SPSS Tools

数据分析实战

基于EXCEL和SPSS系列工具的实践

纪贺元◎著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据分析实战：基于 EXCEL 和 SPSS 系列工具的实践 / 纪贺元著 . —北京：机械工业出版社，2017.4
（数据分析与决策技术丛书）

ISBN 978-7-111-56667-0

I. 数… II. 纪… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2017）第 086855 号

数据分析实战
基于 EXCEL 和 SPSS 系列工具的实践

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）

责任编辑：陈佳媛　　责任校对：殷　虹

印　　刷：中国电影出版社印刷厂　　版　　次：2017 年 5 月第 1 版第 1 次印刷

开　　本：186mm × 240mm　1/16　　印　　张：15（含彩插 0.25 印张）

书　　号：ISBN 978-7-111-56667-0　　定　　价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：（010）88379426　88361066　　投稿热线：（010）88379604

购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

华章IT

HZBOOKS | Information Technology

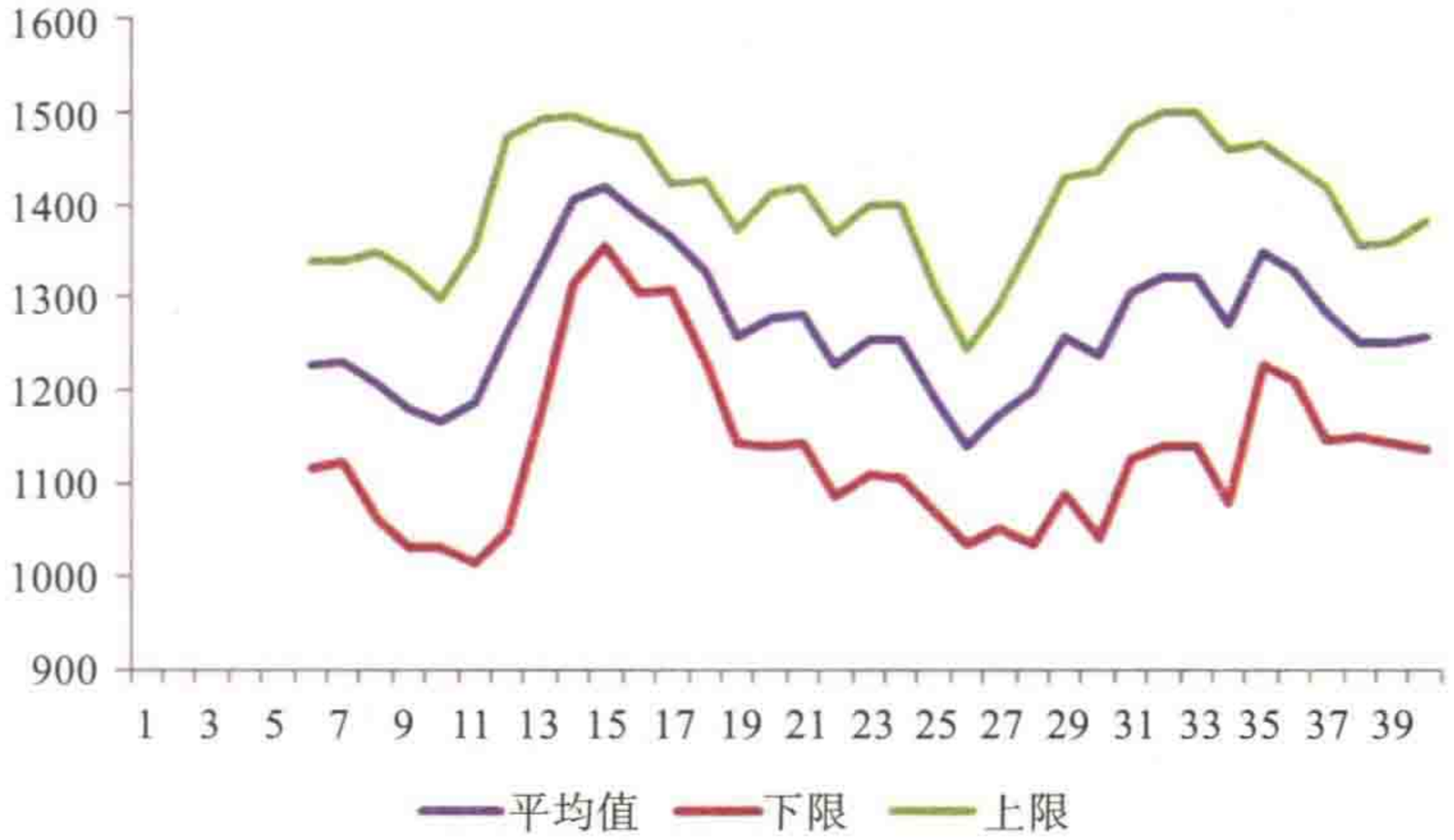


图　2-1

求和项:金额	列标签							
行标签	DVD光驱	打印机	墨盒	内存条	显示器	硬盘	主板	总计
A公司	3.31%	73.23%	5.22%	4.34%	0.00%	8.35%	5.55%	100.00%
B公司	0.00%	0.00%	31.11%	14.20%	42.07%	0.00%	12.62%	100.00%
C公司	11.81%	21.81%	30.56%	2.99%	12.62%	0.00%	20.21%	100.00%
D公司	0.00%	0.00%	14.94%	48.28%	0.00%	36.78%	0.00%	100.00%
E公司	5.57%	39.53%	16.35%	6.51%	24.14%	7.91%	0.00%	100.00%
总计	**5.48%**	**29.47%**	**21.15%**	**9.80%**	**18.91%**	**6.56%**	**8.61%**	**100.00%**

图　5-7

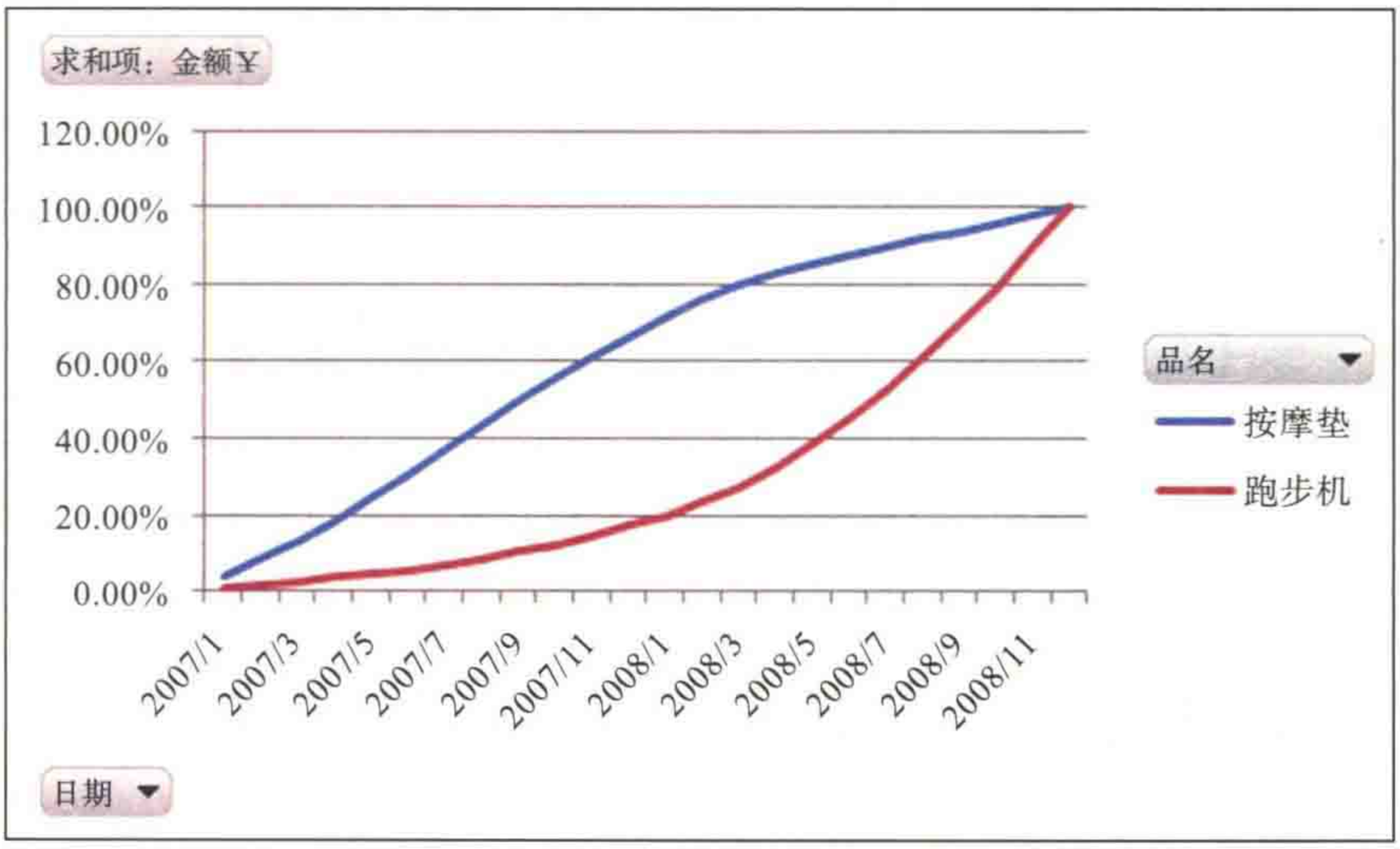


图　5-15

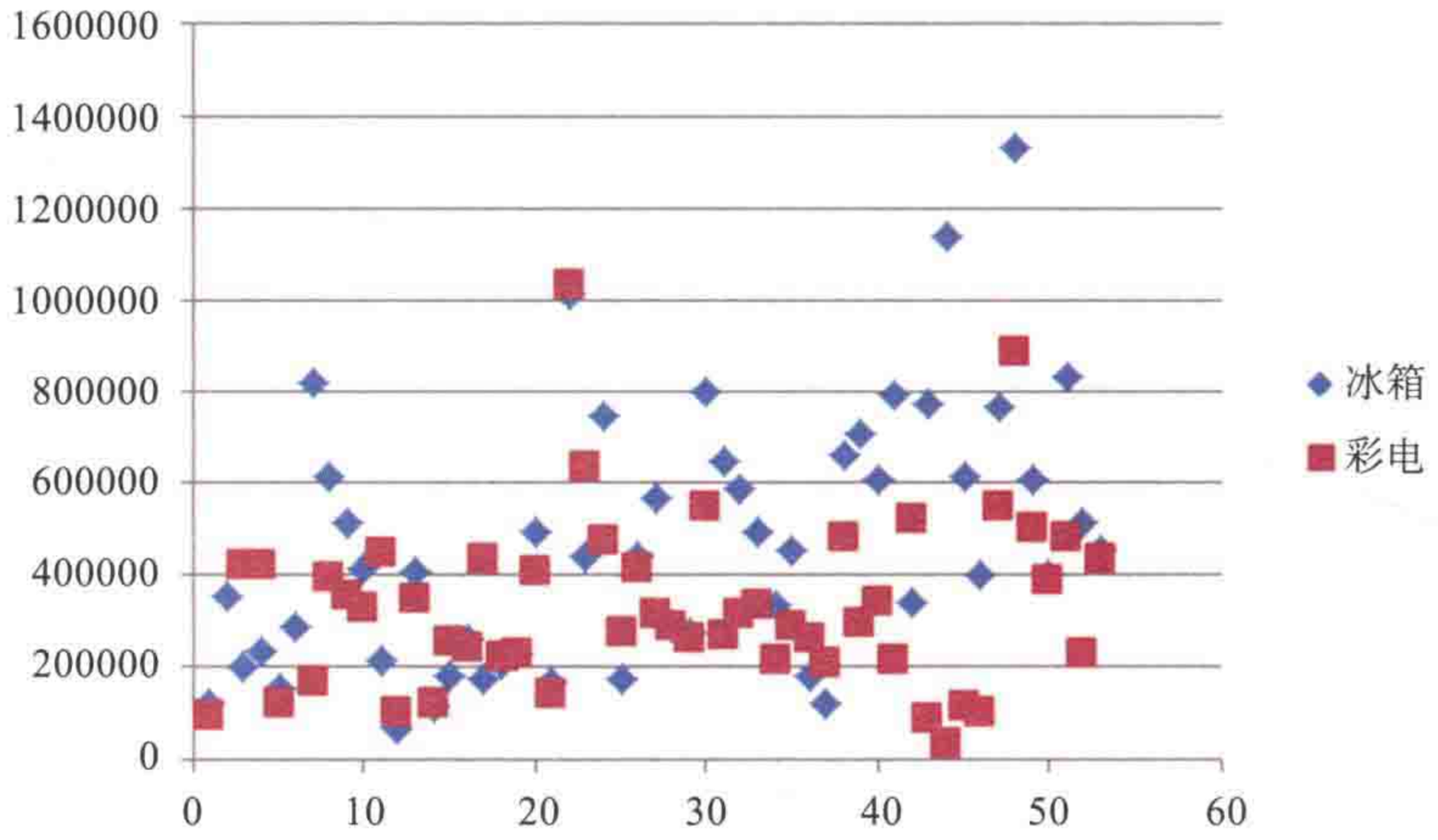

1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
0
10
20
30
40
50
60
冰箱
彩电

图　6-23

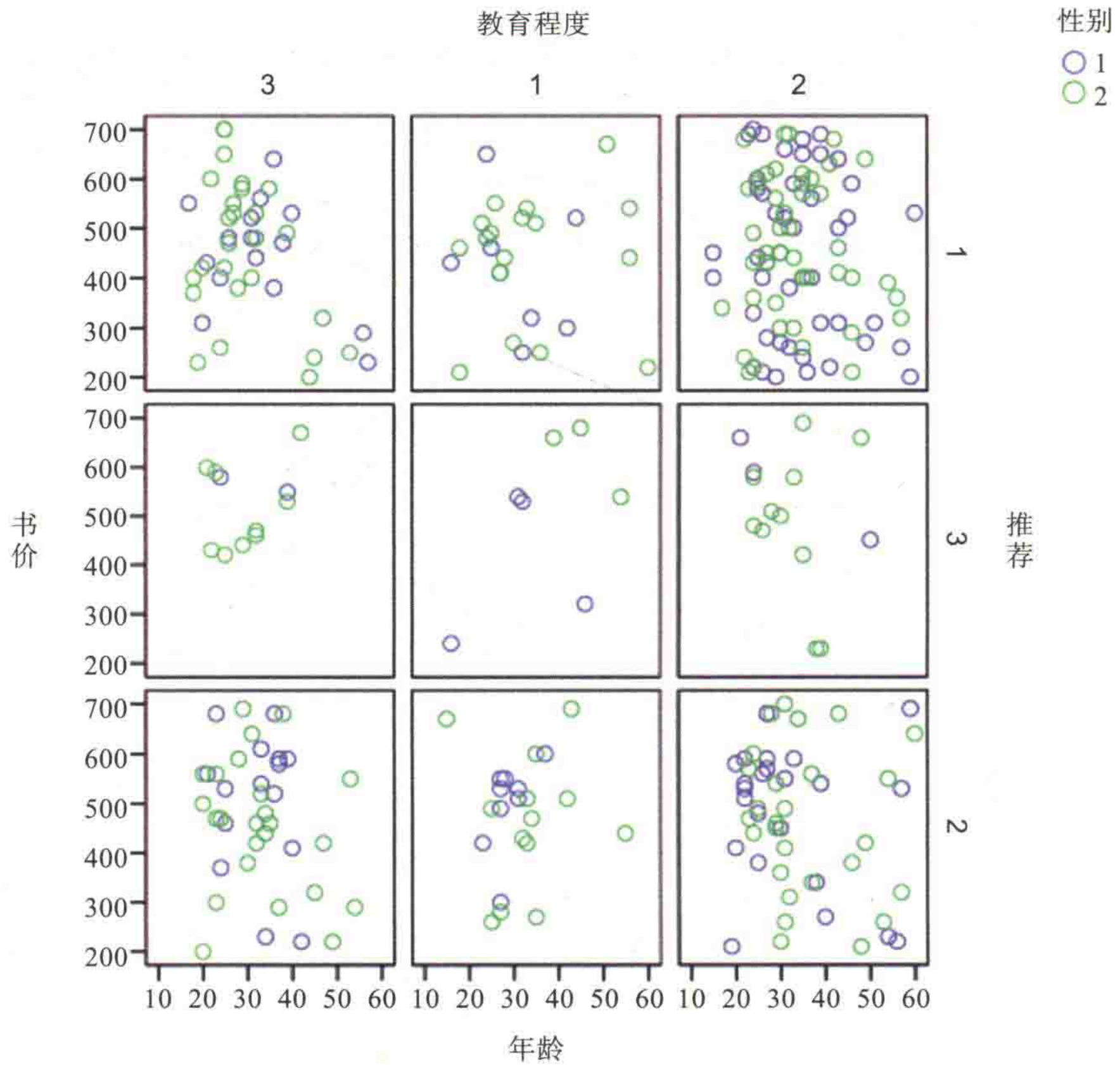

教育程度
性别
1
2
3
1
2
700
600
500
400
300
200
书价
推荐
10 20 30 40 50 60
年龄

图　8-5

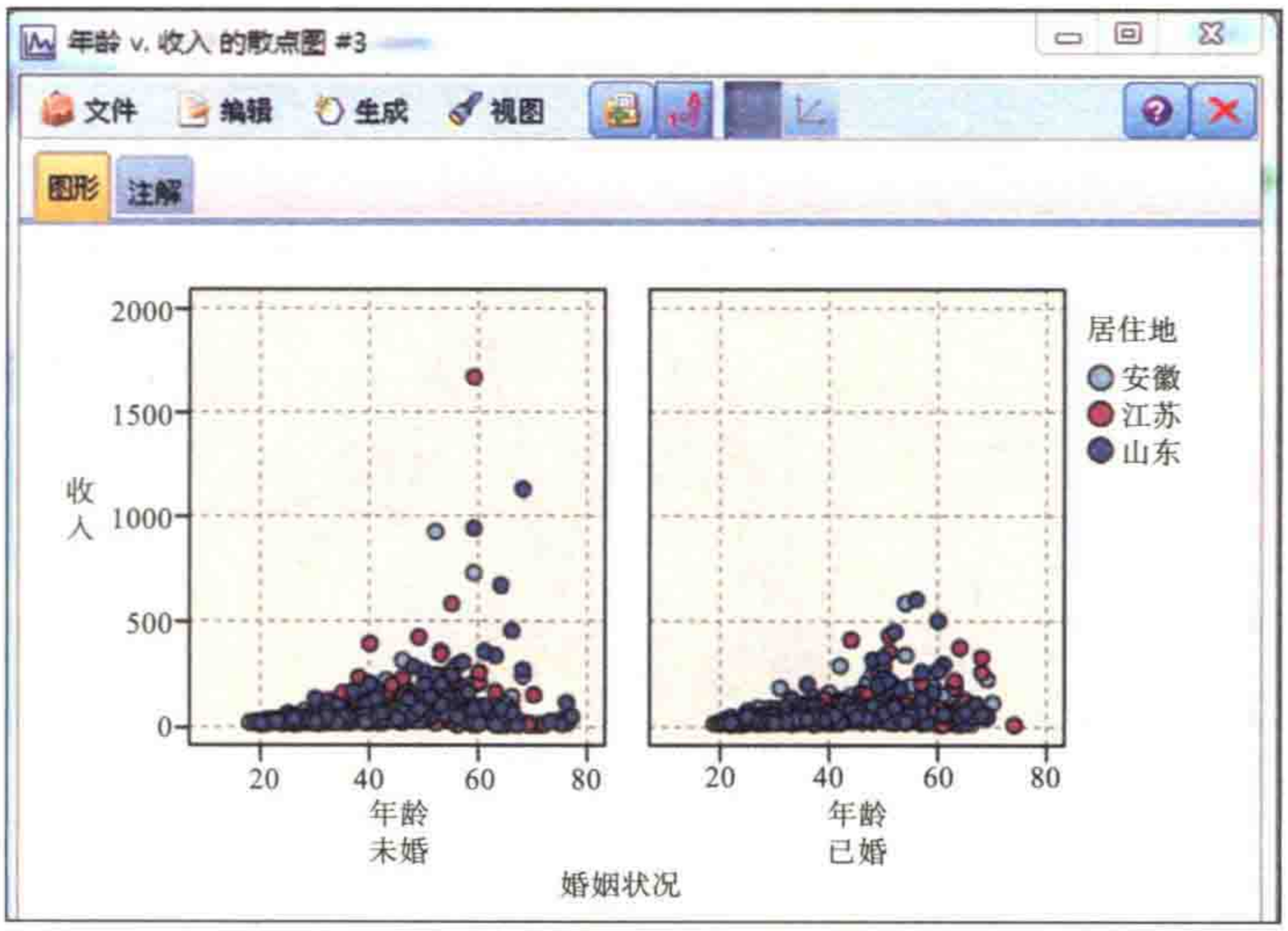


图　8-11

	性别	年龄	因素1	因素2	因素3	因素4	因素5	因素6	因素7	因素8	因素9	因素10	生病
性别	1												
年龄	0.095473	1											
因素1	0.053951	0.203792	1										
因素2	0.031943	0.164544	0.988046	1									
因素3	0.089863	0.20344	0.546053	0.465572	1								
因素4	0.019629	0.2143	0.207408	0.180964	0.233018	1							
因素5	0.014245	-0.06501	0.118708	0.117665	0.095459	0.122085	1						
因素6	0.071829	0.123763	0.375494	0.356265	0.433083	-0.05517	0.101217	1					
因素7	0.17902	-0.24139	-0.07414	-0.06837	0.054606	-0.1062	0.106217	0.193601	1				
因素8	0.225372	-0.03805	0.129716	0.128012	0.118952	-0.02861	0.071038	0.298882	0.726015	1			
因素9	0.063953	0.236611	0.344493	0.312333	0.386864	0.217353	0.116816	0.166439	-0.05379	0.085487	1		
因素10	0.011115	0.14536	0.141668	0.124989	0.260782	0.177808	0.196972	0.139503	0.130457	0.082296	0.576989	1	
生病	0.129069	0.376995	0.565806	0.533458	0.444359	0.382325	0.257899	0.259591	-0.11172	0.019388	0.518459	0.343262	1

图　9-5

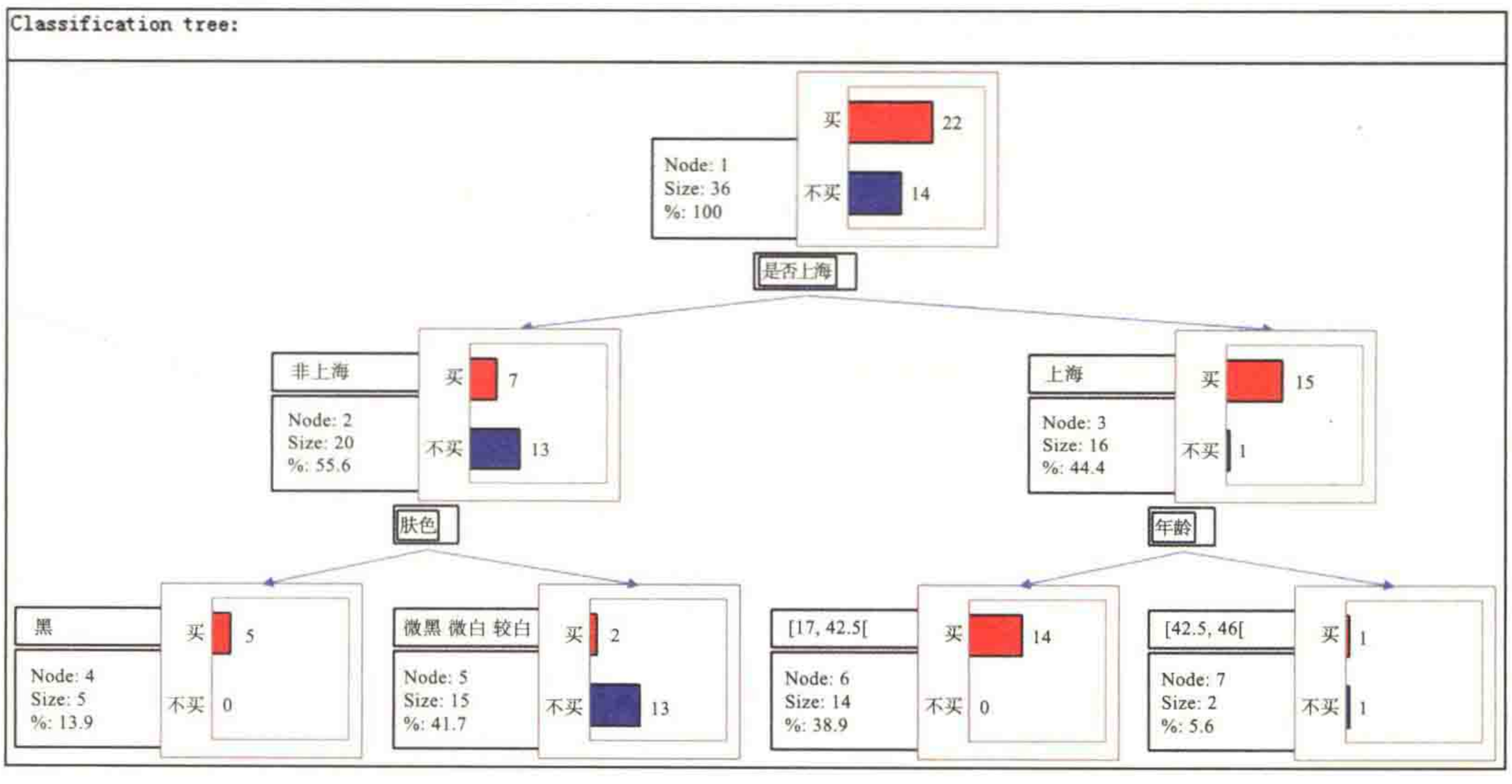


图　9-27

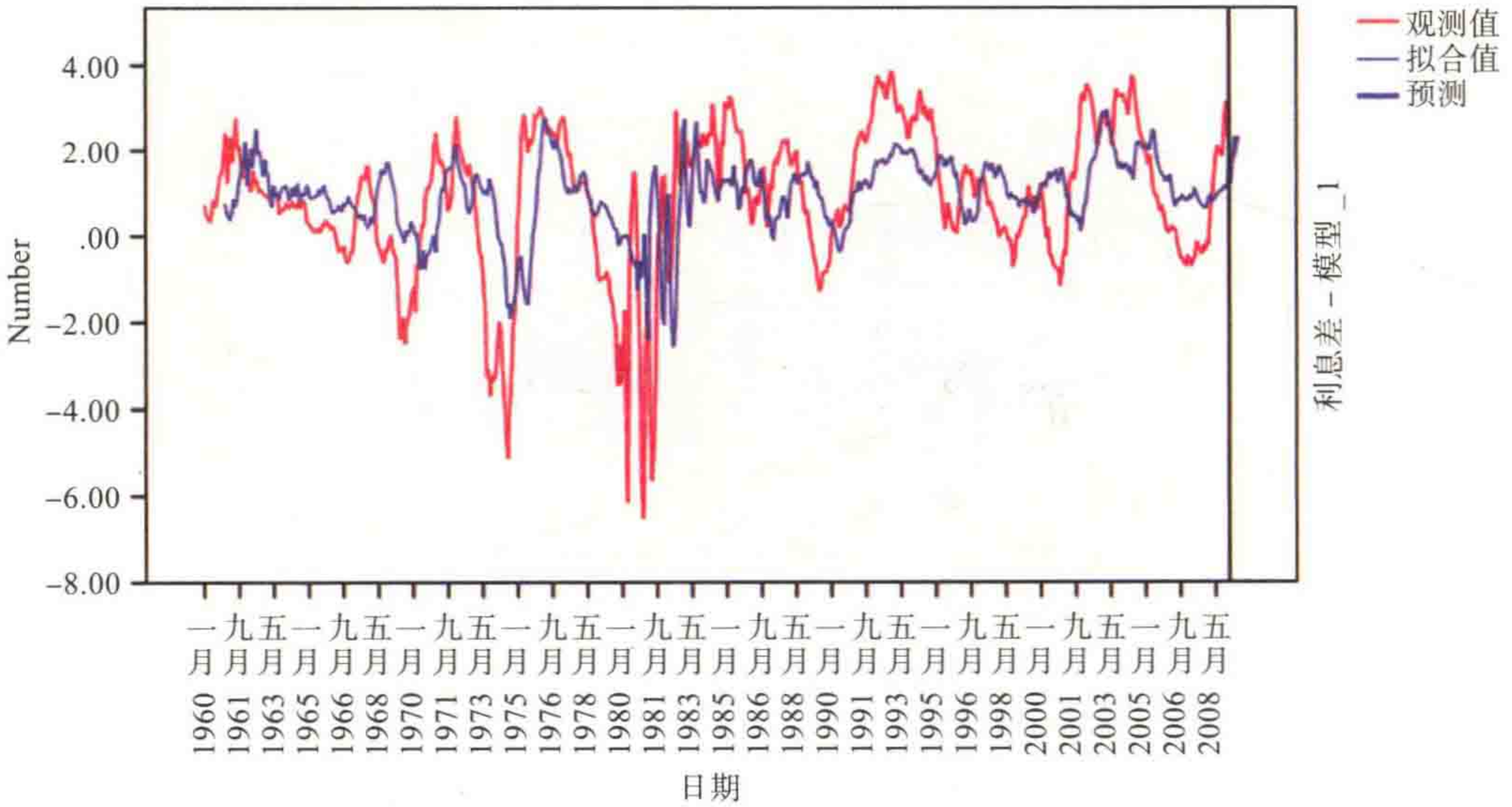

观测值
拟合值
预测
4.00
2.00
.00
-2.00
-4.00
-6.00
-8.00
Number
利息差 - 模型_1
一月 1960
九月 1961
五月 1963
一月 1965
九月 1966
五月 1968
一月 1970
九月 1971
五月 1973
一月 1975
九月 1976
五月 1978
一月 1980
九月 1981
五月 1983
一月 1985
九月 1986
五月 1988
一月 1990
九月 1991
五月 1993
一月 1995
九月 1996
五月 1998
一月 2000
九月 2001
五月 2003
一月 2005
九月 2006
五月 2008
日期

图　13-26

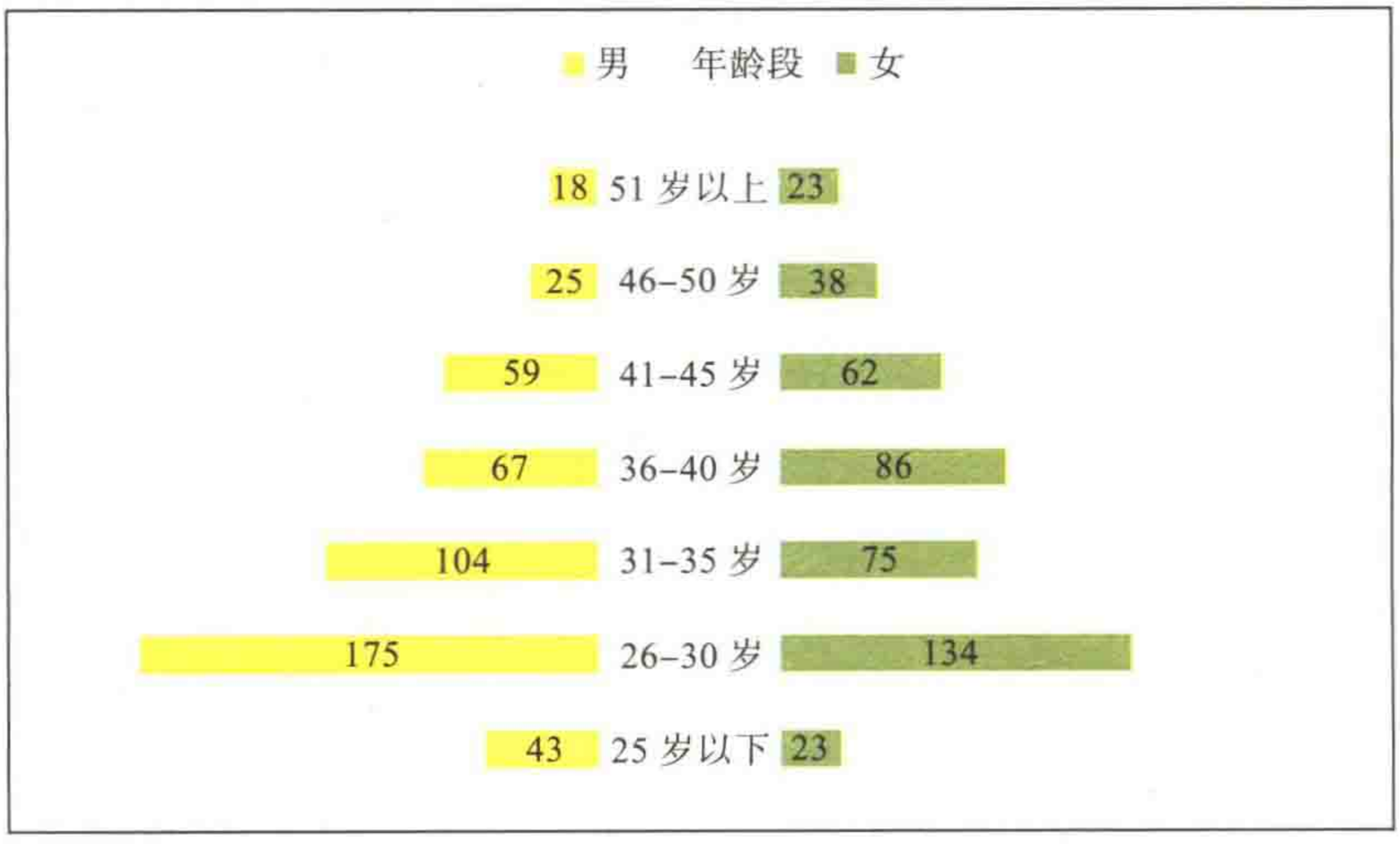

男
年龄段
女
18 51 岁以上 23
25 46–50 岁 38
59 41–45 岁 62
67 36–40 岁 86
104 31–35 岁 75
175 26–30 岁 134
43 25 岁以下 23

图　14-35

Preface 前　言

为什么要写这本书

在我做数据分析培训和咨询的过程中，经常会有学员来问我，有没有合适的统计分析方面的参考书可以推荐。被学员问得多了，慢慢地就有了写本书的冲动，一是毕竟自己写的书和培训的内容比较配套，二是写书对自己来说也是一个总结和提高的过程吧。

“理想很丰满，现实很骨感”，原来觉得自己手里有不少案例，各种工具的使用也算是比较熟练，写起书来应该得心应手，进度也会比较快，但是真到开始动手写作时，才发现并不是那么简单。从框架目录的确定、章节内容的选择、语言风格的打磨，到分析结果截图的选择等，每一个环节都需要细细地思量和斟酌。这本书的写作使我从 2016 年 4 月到 11 月的这段时间非常疲劳，颈椎病也复发了，因为在写书的同时，我的数据分析方面的培训并没有停止。

我在写作本书的时候，给自己规定了几个原则：

一是要实用，要能够解决企业工作中的实际问题。

二是要尽可能地降低读者上手的难度，那种操作非常繁复、需要强大坚实的统计分析理论基础，或者需要编程才能实现的功能，我都没有放在本书中。原因很简单，即使本书讲了那些难度比较大的内容，读者也很难真正应用起来。

三是语言风格尽可能轻松活泼一点，尽量避免很严肃、很晦涩的专业术语，我很难做到“寓教于乐”，但还是尽己所能让本书的阅读轻松一点吧。

在本书的写作过程中，我经常提醒自己这三条原则，并且要求自己遵守它们。

简言之，给读者带来一本“有用的、上手比较容易的、读起来比较轻松的”数据分析书，这就是我写这本书的原则和动力。

读者对象

这本书的读者对象是企事业单位中从事数据分析的非统计专业人士：

- 企业中的市场部相关人员，包括市场分析人员、产品设计和研发人员、销售经理等。
- 企业中的生产部人员，包括生产经理、质量控制经理等。
- 企业中的财务部人员，包括财务总监、财务经理等。
- 企业中其他需要经常和各类数据打交道的管理人员和一般工作人员。

如果读者是高校或者科研院所的教师、学生、科研人员，要从事专业学术论文的撰写或者纵向科研项目的研究工作，不建议你将本书作为主要的阅读和学习的书籍，因为使用的工具、模型、方法都会大相径庭，例如撰写学术论文经常要使用 Eviews、Stata 等专业计量工具，而这些专业计量工具在企业中使用的概率非常低。

如何阅读本书

本书分为三大部分，第一部分基础篇（第 1 章和第 2 章）主要介绍数据分析的概念、术语、方法、模型等，为后续的内容展开奠定基础。

第二部分制表篇（第 3 章到第 5 章）介绍数据的采集、整理以及常用数据报表的制作。

第三部分数据分析篇（第 6 章到第 14 章）占据了本书的大部分篇幅，囊括了常用的、有代表性的、实用的功能，包括数据扫描、数据标注、异常值分析、回归等。

正文中所提“案例文件”为本书的配置案例资料，请通过网络自行下载，下载地址为 http://www.hzbook.com。

勘误和支持

由于作者的水平有限，编写的时间也很仓促，书中难免会出现一些错误或者不准确的地方，恳请读者批评指正。如果你发现本书有错误，或者有其他宝贵意见，请发送邮件到我的邮箱 jhyjhy8888@163.com，我很期待能够收到你们的真挚反馈。

致谢

我跟我的家人说，我这本书是以“part time”的方式写出来的，因为在写书的过程中，我还在四处上课，也做了一些小的咨询项目。

多年以后，如果回顾2016年，我给哪些企业上过哪些课，我未必能记清楚，但是2016年我写作了平生第一本书，这点我不会忘记。

感谢机械工业出版社华章公司的编辑杨绣国老师，感谢你的魄力和远见，在这一年多的时间中始终支持我的写作，你的鼓励和帮助引导我顺利完成了全部书稿。

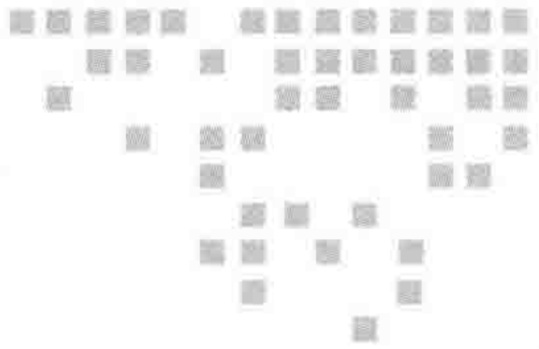
最后我一定要感谢我的家人，是你们给了我一个温暖的港湾，让我在这一年中几乎不用做家务，专心从事培训和本书的写作，多谢多谢！

谨以此书，献给我最亲爱的家人，以及众多热爱数据分析的朋友。

纪贺元

2017年1月于中国上海

目　　录 *Contents*



第1章 Chapter 1

什么是数据分析

1.1 一眼就看到结论还需要数据分析吗

在我做数据分析培训和咨询的时候，时不时会有学员或者客户流露出这样的情绪：

我们的企业其实是不需要数据分析的。

我们公司的业务情况，我很清楚，分析不分析都那样，反正我都知道了。

公司的数据好简单啊，就那么几列，有啥好分析的。

公司里面的很多数据都是造假的，没有分析的价值。

在以上问题中，除了数据质量，其他问题都与企业数据的可分析度有关。数据质量确实是数据分析很难解决的问题，如果企业员工出于种种原因总是在编造各种假数据，这应该属于职业道德或者企业管理水平（企业应该通过严格严谨的管理流程使得员工无从造假）的范畴，这里暂且不讨论。那么，什么是数据的可分析度呢？

这个问题实际上包含如下两层意思：

1）这个企业的数据是比较复杂的，一眼是看不到结论的，需要使用一些工具、模型、方法进行分析。

2）关于数据的分析是有价值的，也就是说分析的过程和结论对于企业是有价值的，能够对企业的生产经营等带来促进和提高。

因此，在数据的可分析度方面，我们需要有一些判断的维度，以帮助我们辨识数据是否值得分析，这里所说的维度主要考虑企业数据量、数据复杂度、数据颗粒度这三个

方面（如图 1-1 所示）。

图 1-1

1.1.1 企业数据量

企业数据量是企业可分析度的第一要素，企业数据量的大小往往取决于两个因素：

一是企业的行业属性，二是企业的信息化程度。众所周知，互联网行业往往也是产生大量数据的行业，“BAT”不仅仅引领了各自行业的发展，同时也是数据行业发展的标杆。

一般情况下，企业的数据量跟企业的规模呈正相关关系，中等以上规模的企业数据量均比较大。但是也有例外，我曾经接触过一家从事智能手机操作系统推送业务的公司，该公司规模很小，只有 40 多人，但是由于合作方是国内诸多智能手机的生产企业，因此该企业的手机用户数量有 3000 多万，每天产生的业务数量高达几 GB。

1.1.2 数据复杂度

如果说数据量相当于数据的行，那么数据复杂度就相当于数据的列。某公司营销部曾给我发来的数据样例，总共的列数加在一起是 12 列。该公司要求分析客户数据，但是涉及客户资料的数据基本上就是客户名称、客户行业（行业数据还是不全的）这两列，客户注册资本、销售收入、雇佣人数都没有，怎么分析？

做过数据分析的人肯定都知道“巧妇难为无米之炊”的苦楚！请想想，你提供的客户数据就是寥寥数列，那要怎么去分析？怎么做文章？

到目前为止，并没有什么明确的指标来度量数据量与数据复杂度，我们很难说每天的数据超过 3 万行就算数据量多，或者说数据超过 30 列就算数据复杂。特别是数据复杂度，这中间还有一个数据相关性的问题：以案例文件 1.1 为例，虽然其中的数据是 3 列，但是用 EXCEL 自带的“数据分析”模块中的“相关分析”进行分析（相关系数的函数，后面会详细讲解），我们发现第二列“销售数量”和第三列“销售额”之间的相关系数是