HZ BOOKS
华章IT

从现实到数据，从数据到看见，从看见到实现！大数据时代的“零起点”教科书。
资深SAS专家系统全面阐述SAS可视化分析技术的理论和实践，结合大量的案例展现SAS可视化分析产品解决商业问题和实施商业项目。

SAS核心技术丛书

可视化分析与SAS实现

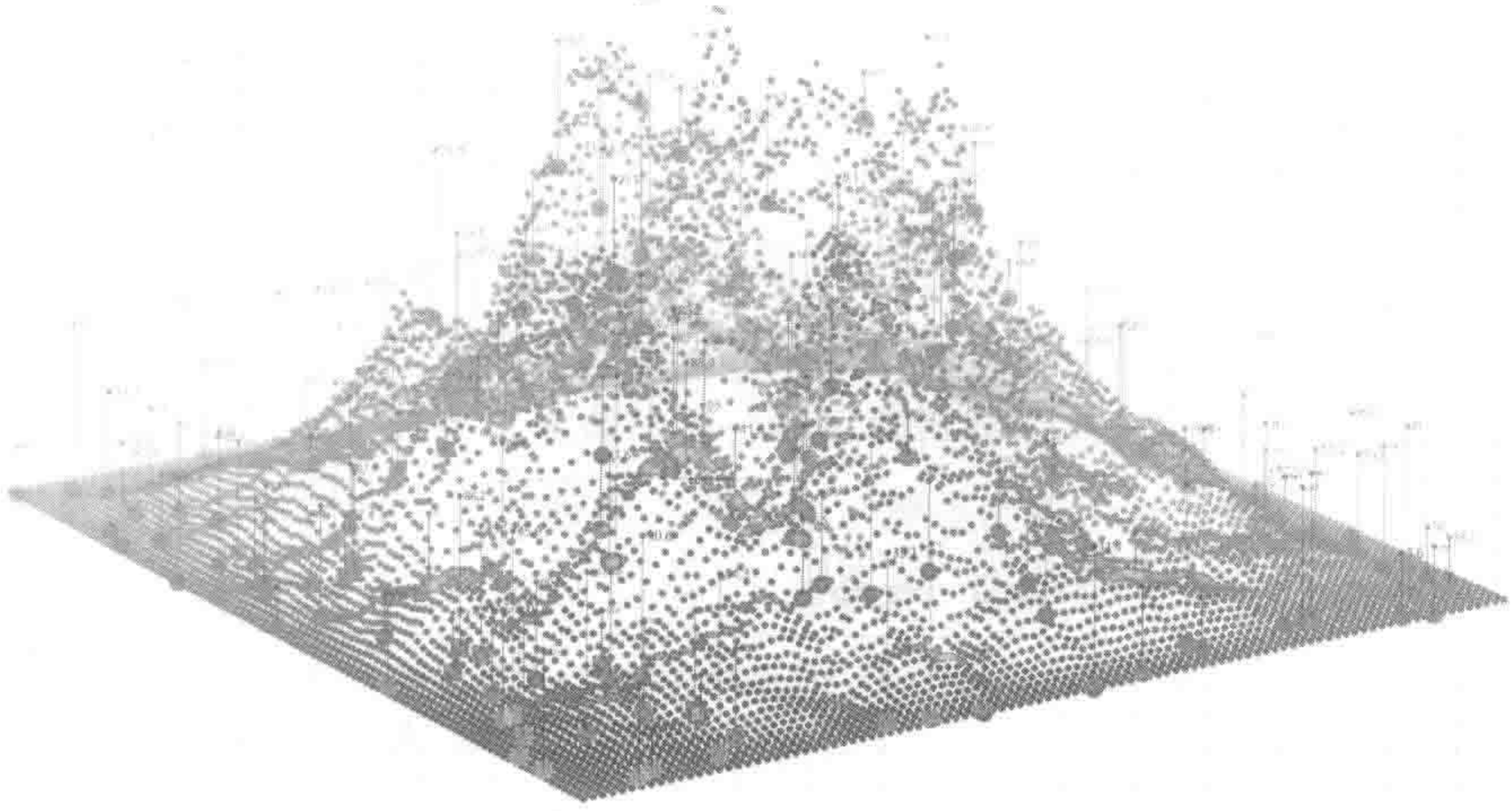
朱继辉 刘政 窦运涛 邱威／著

机械工业出版社
China Machine Press

SAS核心
技术丛书

可视化分析与SAS实现

朱继辉 刘政 窦运涛 邱威／著
夏坤庄／审校



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

可视化分析与 SAS 实现 / 朱继辉等著 . —北京：机械工业出版社，2018.7
（SAS 核心技术丛书）

ISBN 978-7-111-60407-5

I. 可… II. 朱… III. 统计分析 - 应用软件 IV. C819

中国版本图书馆 CIP 数据核字（2018）第 141696 号

可视化分析与 SAS 实现

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：北京市荣盛彩色印刷有限公司　　版　　次：2018 年 7 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：15.5（含彩插 0.25 印张）
书　　号：ISBN 978-7-111-60407-5　　定　　价：69.00 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

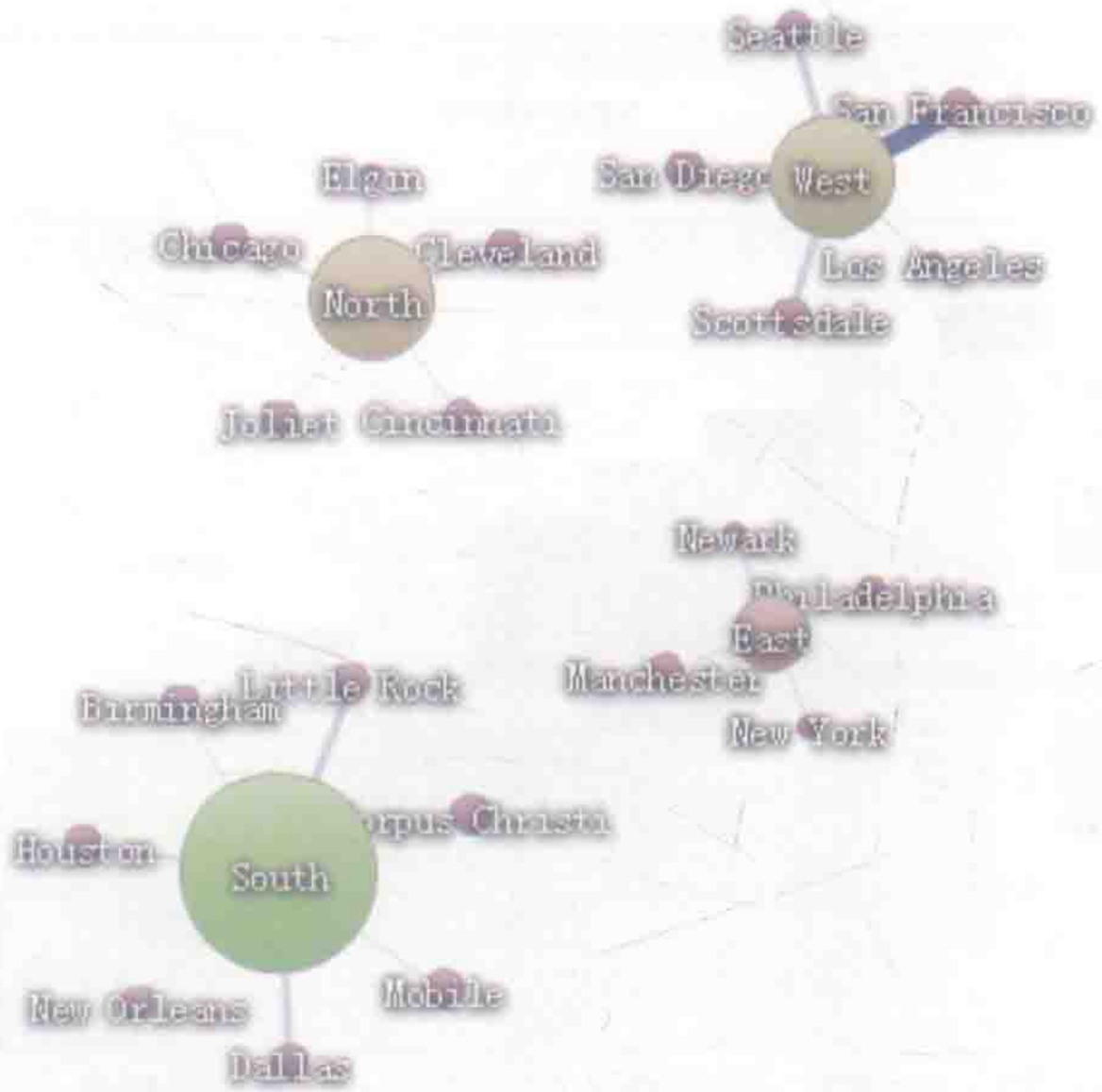


图 2-12　网络图

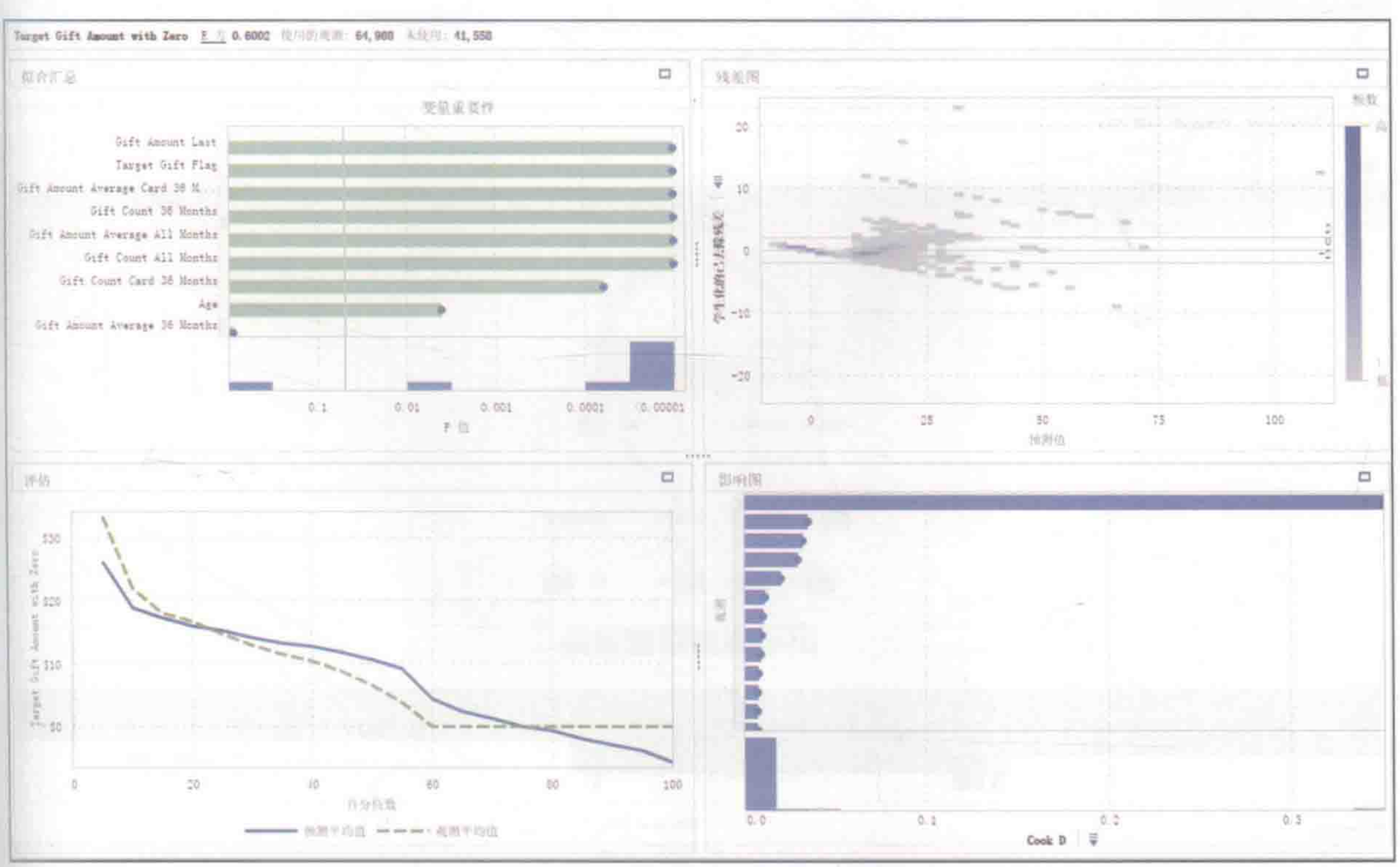


图 6-9　线性回归模型的结果

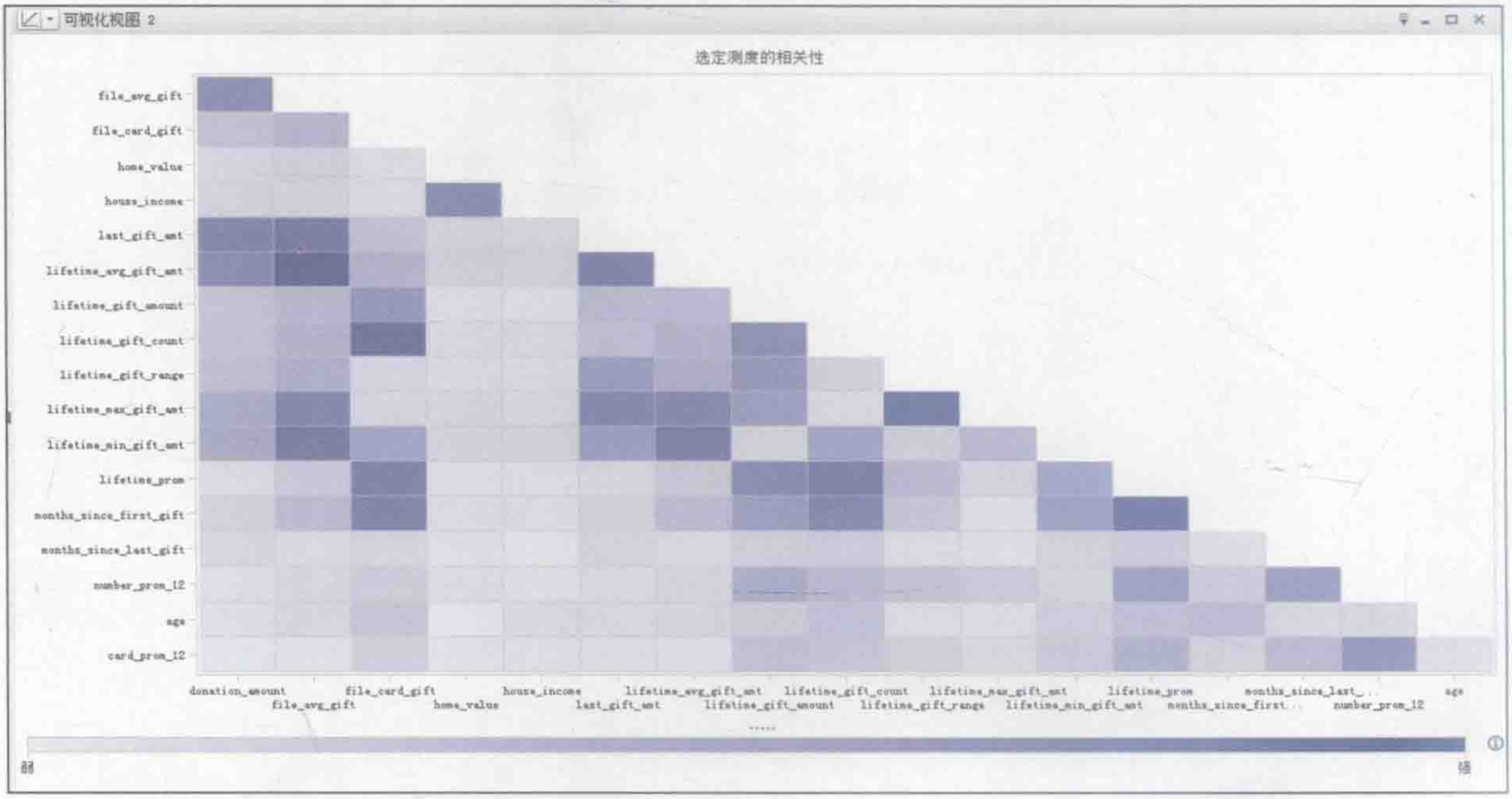


图 6-11　变量相关矩阵图

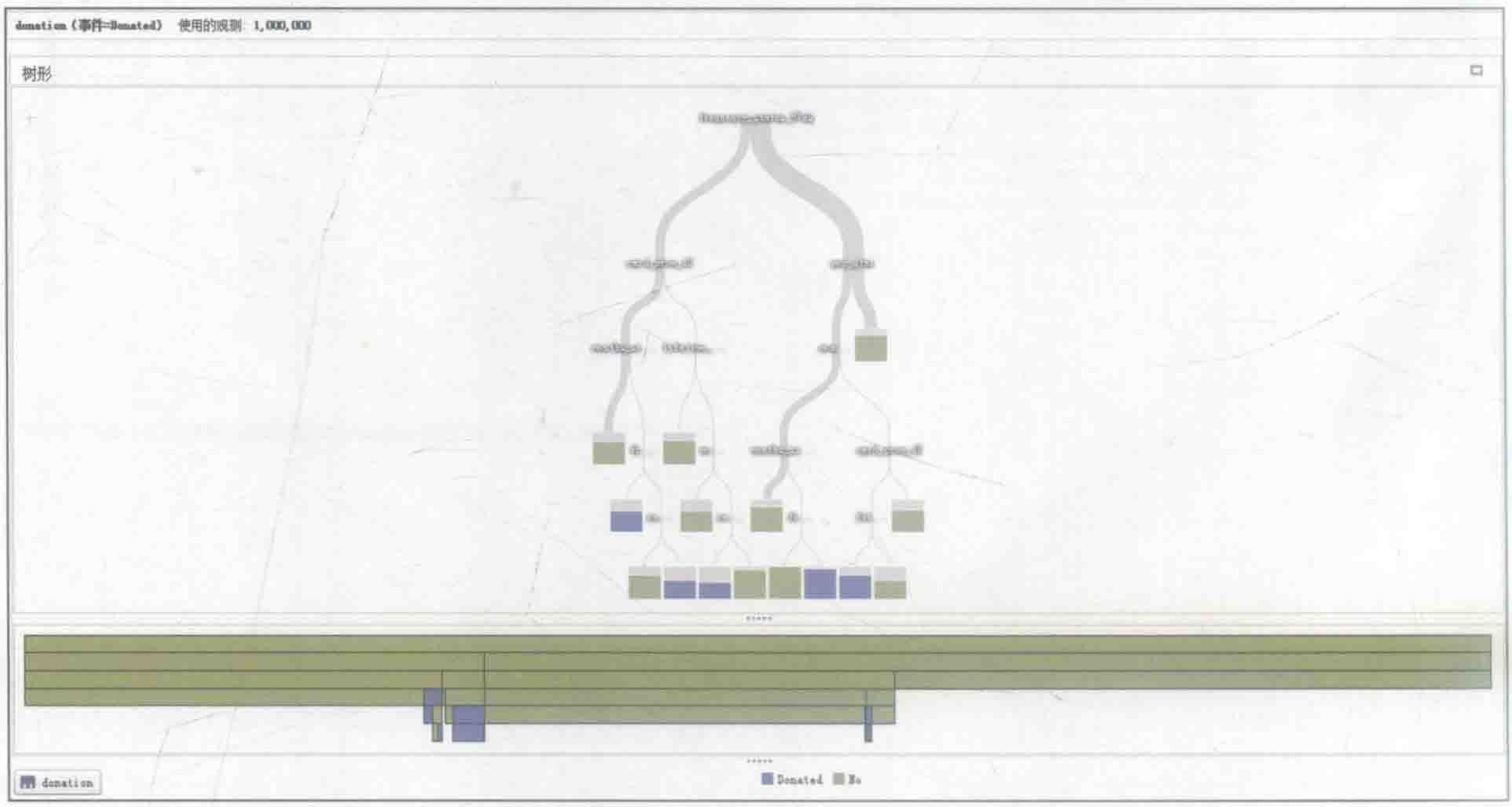


图 6-39　决策树结果

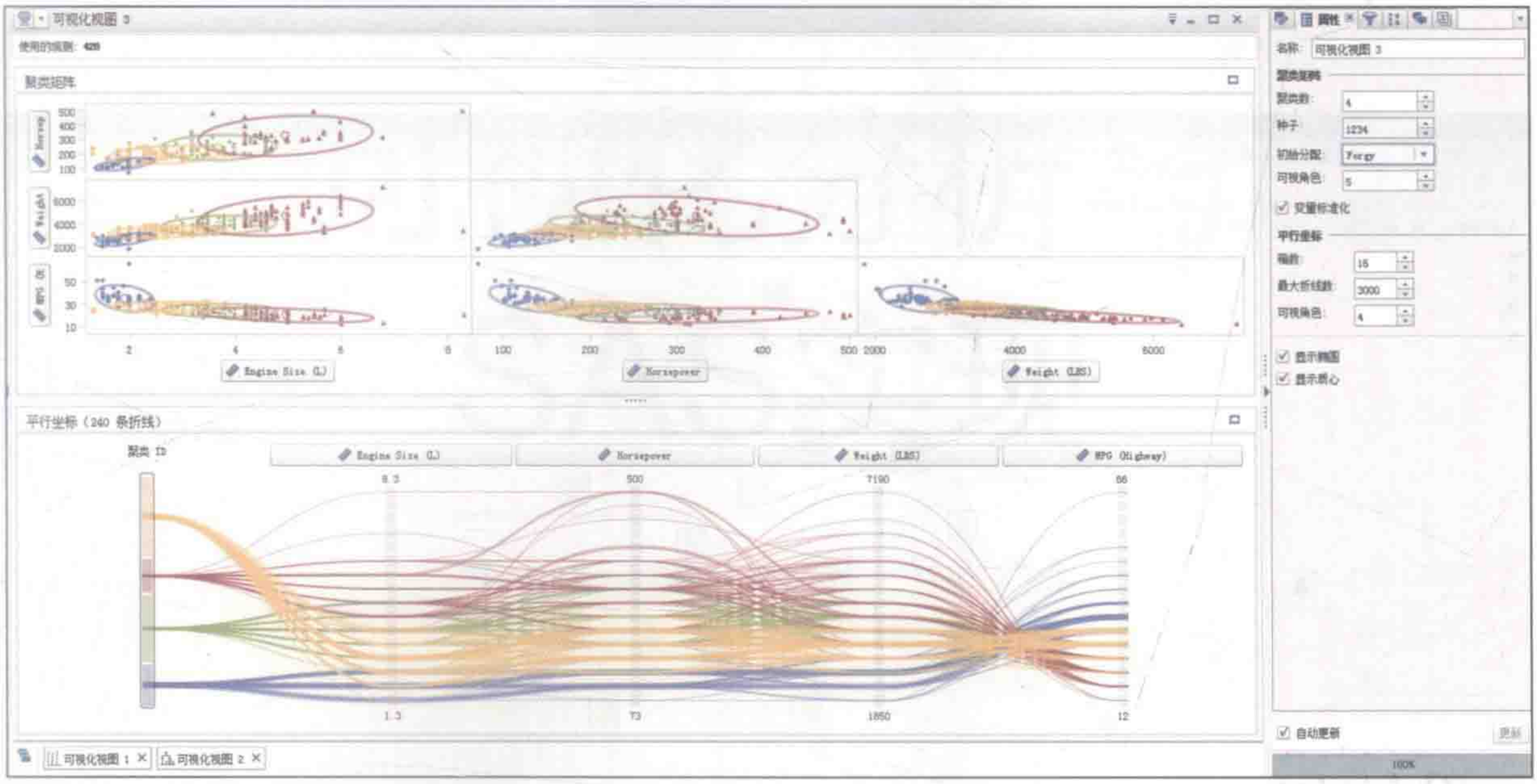


图 6-48　初步聚类结果

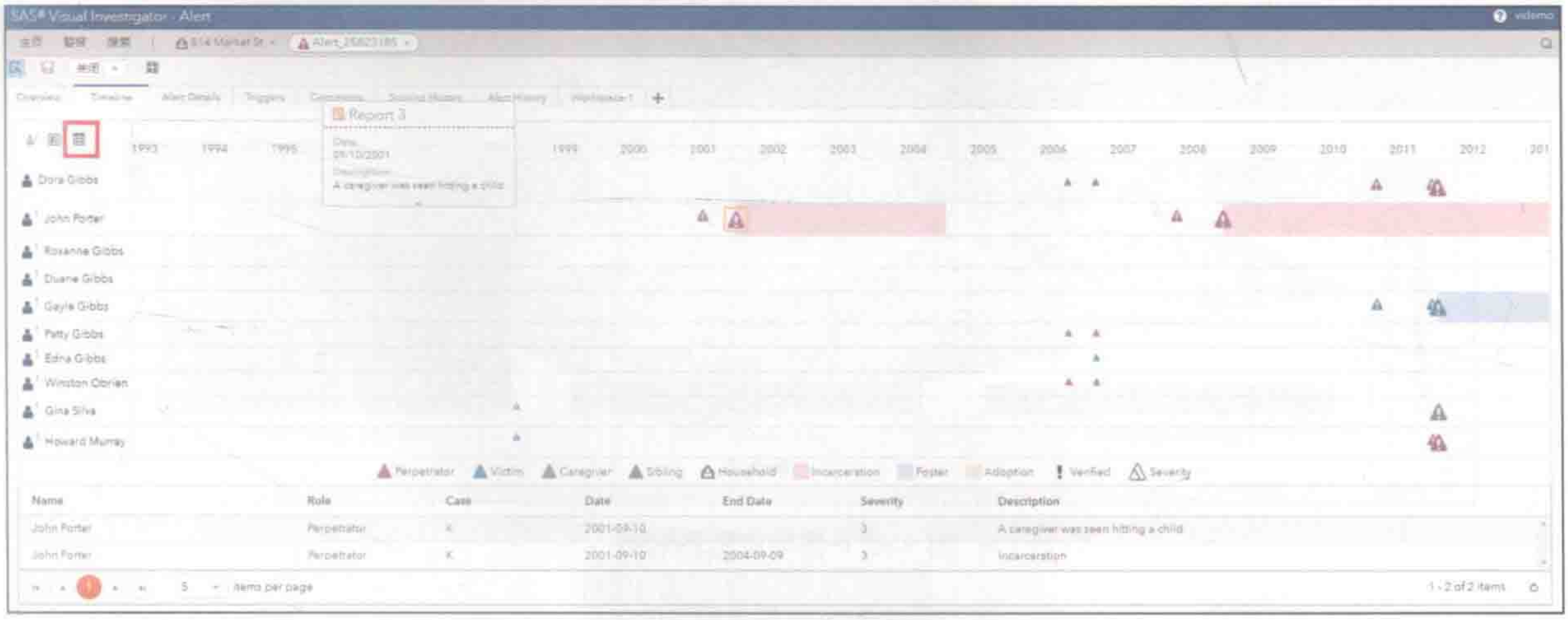


图 7-53　详细表格栏

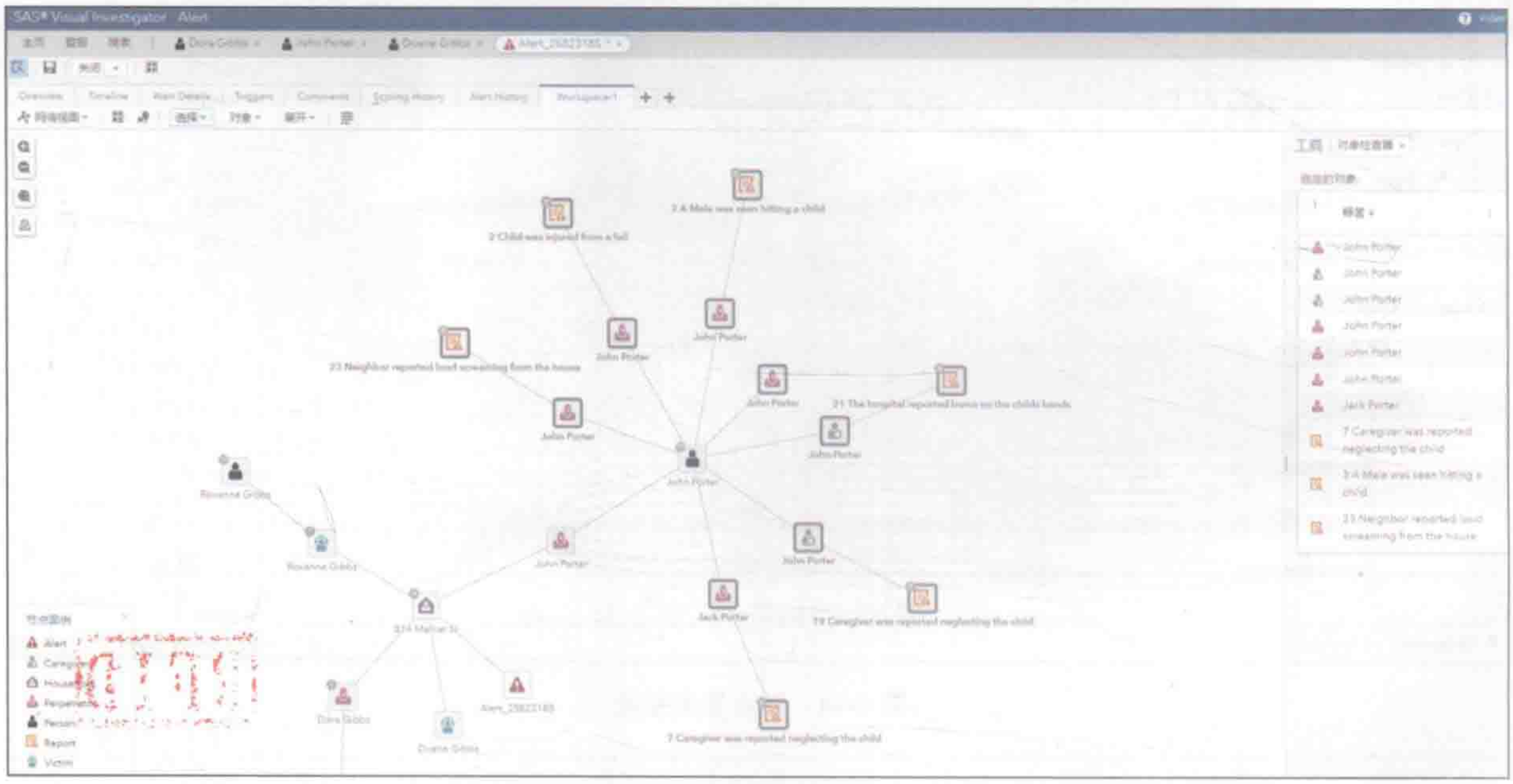


图 7-61　展开 John Porter 的二级网络

图 8-63　进程监控条形图

Preface 前　言

比利时的佛兰芒族地理学家和地图学家亚伯拉罕·奥特柳斯，在研究了一个世纪以来环球旅行探险家们撰写的资料后，于 1570 年在比利时的安特卫普绘制了世界上第一张现代地图集《世界概貌》，即把各种地理上的复杂数据通过图示的方法展示给人们。26 年后，奥特柳斯提出了“大陆漂移学说”的设想。后来，我们还可以在地图上显示人口分布数据，世界宗教分布，世界人民喜欢什么运动的分布，到今天的网民的分布，各国人民喜爱的网站分布，各种调研统计的数据分布，等等。

1812 年夏，俄法战争爆发，拿破仑开始进攻俄国，在战争中遭受了灾难性损失，1813 年以失败告终。法国工程师 Charles Joseph Minard 于 1869 年 11 月 20 日，在巴黎创作完成了一张在信息图界有里程碑地位的“拿破仑 1812—1813 年俄国大进军的人员损失图”。信息图以真实地图为背景，起于波兰 – 俄国边境，止于莫斯科。他在图中使用了 6 个变量的数据：拿破仑军队的数量、行军路线、气温、地理位置、行军到特定地点的时间和距离。线条宽度代表拿破仑的军队人数，黄色表示进攻路线，黑色表示撤退的路线。开始东征时有约 42 万大军，到达莫斯科时剩余约 10 万人，最终返回约 1 万多人。图中下面部分的温度折线图描绘了撤退途中的温度变化，最低温度达到 −37.5 摄氏度。 对比军队规模在撤退途中的阶梯状锐减的转折点与对应的温度变化，排除了当地发生过战役事件后，我们可以直观地推断出撤退时导致士兵死亡的最大原因是气温。

这两张图是数据可视化的经典案例。通过一张图，就把无数的数据汇集在一起，将数据之间的各种联系直观地展示出来，从而揭示出了很多内在的含义。想象一下在当时的环境下，完全用手工的方式把如此多的代表不同维度的数据按照一定的构思汇集在一起，是何等烦琐、艰难、耗时。

20 世纪 60 年代人类就实现了用计算机来做统计分析运算，但是到了 80 年代才实现了计算机的图形化显示、可视化的数据展示。这也仅仅是在数字列表的基础上增加了显示简单图

形的功能。到了 21 世纪，计算机技术和互联网技术获得了长足的发展，各种应用也越趋广泛，特别是电子商务、社交媒体、移动应用和 ERP 的广泛应用，极大地促进了数据的增长，而且数据的种类繁多，非结构化的数据占主要分量，由此对数据分析的能力提出了前所未有的挑战。为了应对这些挑战，人们发明了存储这些数据的平台 Hadoop，处理大量数据的高性能分析技术，开发了新的模型和算法处理非结构化数据，用新的计算机图形学技术与模型来展示它们各种内在的关系。我们可以看到，今天我们对统计分析软件的要求与过去已经有了很大的区别。那么这些区别包括哪些内容呢?

传统的统计分析软件主要是分析结构化的数据，这些数据都是存储在关系数据库、纯文本、Excel 等文件中。今天数据种类以非结构化的数据偏多，而且过去的关系数据库已经无法存储这些数据，无论是存储数据量上，还是数据种类上都无法满足要求。Hadoop 既支持分布式存储，又支持非结构化数据存储。因此，我们新的统计分析软件不但要支持传统的数据存储软件，也要支持 Hadoop。

过去要分析的数据量相对来说都比较小，计算时间基本上是可以接受的（即使需要花费几个小时）。但是，今天的数据量有时候几天都不一定能计算出结果来。这就要求我们在计算技术上有新的突破。SAS 使用了网格分布式计算技术，把计算步骤和数据都分成块，用不同的计算器件，不同的 CPU 多线程地进行计算，然后把结果合起来；用库内分析技术，把对数据的分析计算放到数据库内来进行，减少了对数据的提取和传输过程；用内存分析技术，把由硬盘读取和存放数据的过程改到了由内存读取和存放数据。这三项技术中的任何一项都可以极大地提高数据分析速度，三项技术合而为一，可以获得震撼性的效果，使得实时分析成为可能。过去的数据量小，很容易查看，了解数据属性。要查看今天的数据就要困难许多，我们将这一过程称为数据探索。探索的过程，不仅仅是翻看数据，还要试探性地做一些分析结果的展示，整个的探索过程要流畅，不能有明显的延迟。现在的高性能分析技术完全可以做到。

传统统计分析展示的图表通常都是饼图、直方图、折线图、散点图、柱状图、箱式图、仪表盘等。虽然这些图表也是人们经常会用到的图示，但是如今已经远远不够了。今天的可视化技术还可以展示流程图、衍生分支图、气泡图、矩形树图、面积图、树状图、各种地图、词云、瀑布图、漏斗图、网络结构图等种类繁多的图形，以满足不同的展示和分析需求。

SAS 作为统计分析软件的领导者，早在 2012 年就发布了可视化分析软件“Visual Analytics”，简称 VA。2016 年，SAS 又推出了 Viya，新一代的云上数据分析平台，而 VA 成为所有在 Viya 上运行的行业解决方案的模板。VA 是基于高性能分析技术的，支持 Hadoop，其可视化功能涵盖了整个数据分析的全生命周期，并且简单、易用，给用户带来全新的数据

分析体验。VA 还提供了 21 种可视化视图和分析方法，支持对结构化、半结构化和非结构化数据的可视化分析，支持多用户的信息共享和移动技术。SAS 在高级分析领域占有绝对的领先地位，因此，VA 不仅支持普通商务智能级别的分析，还支持高级分析，就是支持全级别的数据分析，这也是 SAS 可视化分析产品与其他厂家不一样的地方。

本书比较全面地介绍了可视化分析的基本概念、技术组成和产品的架构。通过本书的学习，读者除了可以了解可视化的知识以外，还可以学习可视化分析的基本方法。本书特别适合于那些希望通过简洁、快速的方法就能够进行数据管理，进行数据探索；无须写代码就能进行数学建模；设计各种实用报表方便决策的数据分析人员和相应的管理人员。对于进入数据分析的初级人员，本书也是一本不错的指南。

本书共 8 章。前两章主要介绍可视化分析的基本概念和技术。第 3 ~ 6 章涉及整个数据分析的生命周期。第 3 章介绍数据管理；第 4 章介绍了报表的制作；第 5 章介绍商务智能分析；第 6 章介绍统计分析和数据建模。最后两章是可视化的基本应用。第 7 章介绍可视化反欺诈方面的内容；第 8 章介绍可视化的企业级部署。

本书的完成，来自于整个创作团队的辛勤耕作。大家利用自己的休息时间，一遍一遍地查阅资料，构思内容，完成配图，才使得本书得以和各位读者见面。在这里我要衷心地感谢大家的付出和各位家庭的支持。感谢那些以各种方式为本书的完成提供了帮助的同事和朋友。

SAS 公司在过去的 40 多年里，为行业贡献了各种里程碑式的产品，包括我们在书中要给大家介绍的可视化分析产品。在这里我们要感谢 SAS 公司开发的优秀产品，感谢公司提供的工作学习环境和各种资料，以及对出版本书的支持。

最后，要特别感谢机械工业出版社华章公司的编辑们。感谢他们对于本书出版的指导和帮助。

刘政

2018 年 5 月于北京

目　　录 *Contents*

前言

第1章　可视化分析概论 ····· 1

1.1　可视化分析的意义 ····· 1

1.2　数据可视化分析兴起的背景 ····· 3

1.3　数据分析的可视化与分析的不同层次 ····· 4

1.3.1　数据获取与数据转换 ····· 4

1.3.2　高级分析与模型开发 ····· 5

1.3.3　分析结果展现与模型应用 ····· 5

1.4　可视化分析面临的挑战与应对 ····· 6

1.4.1　可视化分析面临的挑战 ····· 6

1.4.2　SAS 的可视化分析实现 ····· 7

1.5　本章小结 ····· 9

第2章　SAS可视化分析技术概述 ····· 10

2.1　SAS 数据可视化分析的平台基础 ····· 10

2.2　SAS 可视化分析家族成员、主要功能和相互联系 ····· 11

2.2.1　SAS 可视化分析 ····· 12

2.2.2　SAS 可视化统计 ····· 13

2.2.3　SAS 可视化调查 ····· 14

2.2.4　SAS 可视化数据挖掘和机器学习 ····· 15

2.2.5　相互联系 ····· 16

2.3　SAS 可视化分析功能概述 ····· 17

2.3.1　数据导入 ····· 17

2.3.2　数据处理 ····· 18

2.3.3　数据分析 ····· 18

2.3.4　基于 Web 的报表设计 ····· 22

2.4　SAS 数据可视化分析的展望 ····· 24

2.5　本书内容概述 ····· 24

2.6　本章小结 ····· 24

第3章　SAS Visual Analytics的数据访问和准备 ····· 25

3.1　认识数据源 ····· 26

3.1.1　单一文件类型 ····· 26

3.1.2　数据库和大数据存储 ····· 27

3.2　使用 Administrator 管理 LASR 服务器 ····· 27

3.2.1　创建 LASR 服务器 ····· 29

3.2.2　创建并配置 HDFS 目录 ····· 32

3.2.3 启动 LASR 服务器 …… 33
3.2.4 加载单一 SAS 数据集 …… 34
3.2.5 加载 HDFS 数据 …… 35
3.2.6 高级数据管理 …… 38
3.3 准备数据的最佳实践 …… 39
3.4 如何使用 Visual Data Builder 准备数据 …… 42
3.4.1 使用 Visual Data Builder 的场景 …… 42
3.4.2 导入数据 …… 43
3.4.3 表查询和表连接 …… 46
3.4.4 导入 Information Map 查询 …… 50
3.4.5 追加表 …… 52
3.4.6 创建 LASR 星型表 …… 53
3.5 本章小结 …… 56

第4章 标准报表与定制化报表分析 …… 57

4.1 SAS Visual Analytics Designer 和 Visual Analytics Graph Builder 介绍 …… 57
4.1.1 SAS Visual Analytics Designer …… 57
4.1.2 SAS Visual Analytics Graph Builder …… 57
4.2 创建标准报表 …… 58
4.2.1 使用各类报表对象 …… 58
4.2.2 在 SAS Visual Analytics Designer 中处理和分析数据 …… 58
4.2.3 报表过滤，报表交互，报表链接 …… 62
4.2.4 使用报表中的参数 …… 68
4.3 创建定制化报表 …… 71
4.4 共享报表 …… 72
4.5 本章小结 …… 74

第5章 钻取查询与仪表盘 …… 75

5.1 创建钻取查询报表 …… 76
5.1.1 创建层次，生成钻取查询报表 …… 76
5.1.2 编辑层次，更新钻取查询报表 …… 80
5.1.3 创建时间层次，生成钻取查询报表 …… 82
5.1.4 从可视化图形中创建层次 …… 82
5.2 创建仪表盘 …… 83
5.3 本章小结 …… 84

第6章 可视化统计分析与预测模型 …… 85

6.1 SAS Visual Statistics 介绍 …… 85
6.2 SAS Visual Statistics 用户界面以及架构 …… 86
6.3 探索性数据分析 …… 87
6.3.1 探索性数据分析简介 …… 87
6.3.2 SAS Visual Statistics 实现探索性数据分析 …… 88
6.4 线性回归模型 …… 90
6.4.1 线性回归模型简介 …… 91
6.4.2 SAS Visual Statistics 线性回归可视化分析 …… 92
6.4.3 SAS Visual Statistics 线性回归模型举例 …… 95
6.5 逻辑回归 …… 101

6.5.1 逻辑回归模型简介 …… 101
6.5.2 SAS Visual Statistics 逻辑回归可视化分析 …… 103
6.5.3 SAS Visual Statistics 逻辑回归模型举例 …… 104
6.6 广义线性模型 …… 107
6.6.1 广义线性模型简介 …… 107
6.6.2 SAS Visual Statistics 广义线性模型可视化分析 …… 107
6.6.3 SAS Visual Statistics 广义线性模型举例 …… 108
6.7 决策树 …… 110
6.7.1 决策树模型简介 …… 110
6.7.2 SAS Visual Statistics 决策树可视化分析 …… 112
6.7.3 SAS Visual Statistics 决策树模型举例 …… 113
6.8 聚类 …… 116
6.8.1 聚类分析简介 …… 116
6.8.2 SAS Visual Statistics 聚类可视化分析 …… 117
6.8.3 SAS Visual Statistics 聚类分析举例 …… 117
6.9 模型比较和模型评分 …… 120
6.9.1 模型比较 …… 120
6.9.2 模型比较可视化界面 …… 120
6.9.3 模型评分 …… 122
6.10 本章小结 …… 123

第7章 可视化调查 …… 124

7.1 SAS Visual Investigator 介绍 …… 124
7.2 SAS Visual Investigator 的主要功能和系统架构 …… 125
7.3 SAS Visual Investigator 在预防违规或犯罪领域的应用 …… 126
7.3.1 从警报管理中发现高风险活动 …… 127
7.3.2 通过实体分析发现风险活动的诱因 …… 129
7.4 SAS Visual Investigator 在金融欺诈及反洗钱领域的应用 …… 136
7.4.1 生成警报信息 …… 137
7.4.2 在警报控制台中发现风险 …… 140
7.4.3 搜索实体并进行初步调查 …… 141
7.4.4 在工作区中进行详尽调查 …… 142
7.4.5 使用时间滑块进行深度挖掘 …… 147
7.5 SAS Visual Investigator 在法律、政府和社会管理方面的应用 …… 151
7.5.1 基于汇总报告评估风险 …… 152
7.5.2 持续的个案监控 …… 155
7.6 本章小结 …… 159

第8章 SAS可视化分析技术的企业级部署和应用 …… 160

8.1 企业级部署 …… 160
8.1.1 架构设计 …… 161
8.1.2 大规模并行处理部署要点 …… 175
8.1.3 后配置、验证、调优 …… 186
8.2 企业级应用的管理和安全 …… 199
8.2.1 管理概述 …… 199
8.2.2 操作计算服务器 …… 215
8.2.3 环境监控 …… 218
8.2.4 安全 …… 224
8.3 本章小结 …… 236

第1章 Chapter 1

可视化分析概论

1.1 可视化分析的意义

数据可视化分析是通过友好的交互式图形界面，来辅助用户对数据进行复杂处理和分析的科学与技术。数据分析的可视化至少包含两个方面的含义，其一是指在数据分析的过程中，通过直观的图形化界面以交互的方式采用合适的数据分析方法，对复杂的数据进行有效的处理和分析，其二是指在各个分析阶段的分析结果处理中，通过直观的图形化界面以交互的方式采用包括图像在内的多种形式表达展示和传递分享分析的结果。

数据分析的意义在于从数据中发现有意义的信息。可视化数据分析的意义在于让分析的过程更简单直观，让分析的结果更简洁清楚，从而让更多的人可以利用复杂的分析方法来洞察数据，让更多的人可以利用数据分析的结果指导和帮助自己的工作。

如上所述，数据分析的可视化，既体现在通过图形的方式清晰有效地表达和传递信息，也体现在帮助理解和分析复杂的数据。例如，通过数据可视化分析，我们可以将一个包含多个维度信息的数据通过图形化操作界面方便地转化成为用户可以直观查看，并且可以快速解读的图形，这样数据当中蕴含的信息才可以被快速直观地理解，进而使用户可以基于数据中的信息进行有效的决策。

接下来我们通过一个具体的例子展现可视化在数据分析中的作用。首先查看下面的数据集，该数据集有 11 个观测和 8 个变量，见图 1-1。

对数据的描述性统计量进行计算显示，数据中 x1，x2，x3，x4 的平均值均为 54，y1，y2，y3，y4 的平均值均为 37.5，同时 x1，x2，x3，x4 的方差均为 396，而 y1，y2，y3，y4 的方差也很接近，在 103 左右（如图 1-2 所示）。

	x1	x2	x3	x4	y1	y2	y3	y4
1	60	60	60	48	40.2	45.7	37.3	32.9
2	48	48	48	48	34.75	40.7	33.85	28.8
3	78	78	78	48	37.9	43.7	63.7	38.55
4	54	54	54	48	44.05	43.85	35.55	44.2
5	66	66	66	48	41.65	46.3	39.05	42.35
6	84	84	84	48	49.8	40.5	44.2	35.2
7	36	36	36	48	36.2	30.65	30.4	26.25
8	24	24	24	114	21.3	15.5	26.95	62.5
9	72	72	72	48	54.2	45.65	40.75	27.8
10	42	42	42	48	24.1	36.3	32.1	39.55
11	30	30	30	48	28.4	23.7	28.65	34.45

图 1-1　数据集列表

Variable	Mean	Variance
x1	54.00	396.00
x2	54.00	396.00
x3	54.00	396.00
x4	54.00	396.00
y1	37.50	103.18
y2	37.50	103.19
y3	37.50	103.07
y4	37.50	103.08

图 1-2　数据集变量描述统计量

通过计算数据集当中 4 对变量（x1，y1），（x2，y2），（x3，y3），（x4，y4）的相关性，发现相关系数均为 0.816。

如果只对数据集当中 4 对变量的均值、方差以及相关性计算数值，而不进行数据可视化分析，除非分析者具备比较全面的统计学知识和思维习惯，否则也许会得出这样的结论：4 对变量的关系是一样的。可是当我们尝试将 4 对变量分别进行可视化分析，用数据集当中的 11 个观测生成散点图时，我们就会得到图 1-3 所示的结果。

这时候，我们不难发现 4 对变量之间的关系存在较大差异。也就是说虽然 4 对变量在均值、方差、相关性上都一致，但是可视化分析显示了它们各自之间的特殊关系。可以看到在（x3，y3）和（x4，y4）的散点图中显著存在的离群值，同时（x2，y2）的关系不是简单的线性关系。这个例子简单印证了数据可视化分析在揭示数据之间隐藏关系方面所具有的重要作用。一般来说，数据可视化分析的益处可以归纳为以下几个方面：

- 数据可视化分析使得数据中所蕴含的信息更直观，更容易被理解，同时数据可视化分析还可以发现数据之间隐藏的关系。
- 数据可视化分析使得数据分析的门槛降低，业务人员可以通过可视化分析界面去获取数据，探索数据，进行数据分析。
- 数据可视化分析可以让用户更容易和数据进行交互，数据可视化分析赋予了业务人员新的“语言”，使他们可以更有力地利用数据去表达观点。

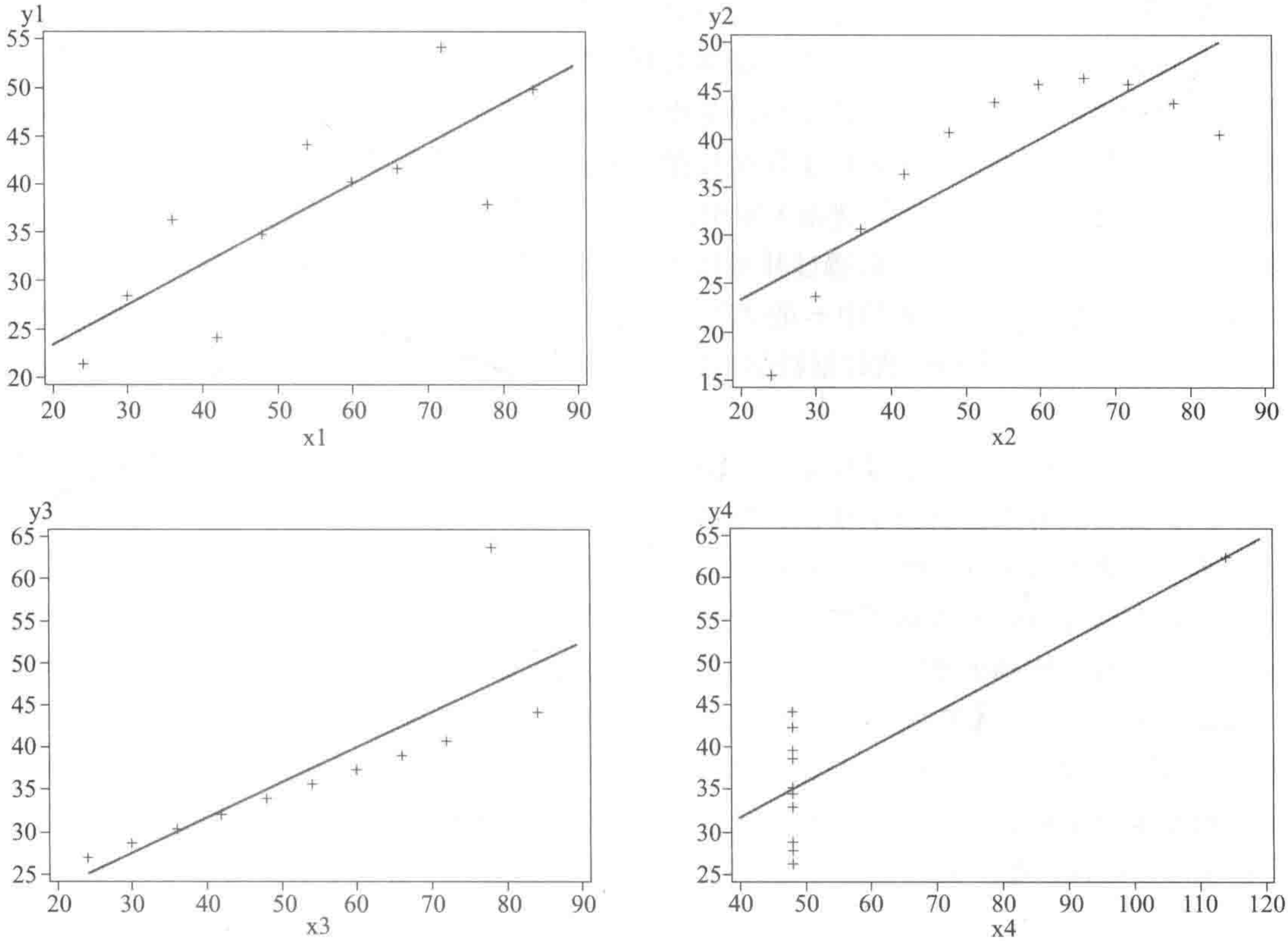


图 1-3　变量关系可视化展现

1.2　数据可视化分析兴起的背景

上面的例子简单介绍了可视化分析如何帮助更好地理解和分析数据。但是，很显然，仅凭上面提到的技术性优势是不可能让这一领域成为数据分析行业的一个热点的。那么数据可视化分析为什么会变得如此重要？究其主要原因，还是由于数据分析日趋重要而引起的对数据分析需求的不断增长和对高素质数据分析人员的巨大需求。

数据已经成为各个组织机构的宝贵资产，如何有效地利用数据了解过去、管理现在、预测并且优化未来成为它们发展的重要问题，数据分析已经成为提升企业竞争力的关键环节。越来越多的组织机构依靠正确可靠的信息来进行决策并取得成功，而其中绝大部分正确可靠的信息出自数据分析，所以在这些组织机构中，逐渐出现了数据科学家的角色，并且这个角色显得日益重要。

各机构对于数据科学家的期望是能够对海量的数据进行处理，并采用适当的算法从海量数据当中获取有价值的信息。具体来说，数据科学家的职责体现在数据价值链的四个阶

段：数据产生、数据获取、数据存储和管理、数据分析。如果把数据当作是原始资料，前两个阶段是资料采集阶段，而数据存储和管理与数据分析则是对这些原始资料进行深加工产生巨大价值的阶段。由于这四个阶段所需要的技能各不相同，所以一名出色的数据科学家也需要掌握应对各个不同阶段工作的技能。具体来说，数据科学家需要有一定的数学知识，尤其是统计学和矩阵运算的相关知识；另外，数据科学家应该有较强的程序开发能力，能够对算法和处理数据的逻辑通过开发代码实现；其次，数据科学家需要具备快速理解业务背景和问题的能力，在现实中不难发现，很多数据科学家也是某个领域（例如金融或供应链等领域）的业务专家；当然数据科学家还应当善于沟通，善于将分析的过程和分析的结果和别人分享。

对数据科学家的这些要求和他们所需要承担的责任，使得寻找合适的数据科学家并非易事。事实上具有丰富的数学知识，高超的编程经验，并且具有相当的行业领域知识的人才是非常稀缺的。而对于数据科学家的需求则是不断增加的。由此就带来两个问题，第一，如何降低数据分析的工作强度以使数据科学家能够承担更多的工作？第二，如何采用有效的技术与工具，使得更多的人可以分担数据科学家的工作？数据可视化分析技术就是在这样的背景出现并飞速发展的。好的数据可视化分析工具为具备一定业务知识以及数学知识，但对于计算机程序开发了解较少的人才提供了对大量数据进行快速有效分析的利器。可视化分析技术提供的自助式的数据准备、数据转换，交互式的数据探索以及容易上手的高级分析技术，可以让更多的人员经过短期的培训就能够处理和分析大量的数据。

1.3 数据分析的可视化与分析的不同层次

数据分析的可视化是指数据分析过程的可视化和数据分析结果的可视化。一个完整的数据分析过程包括数据获取、数据的清洗与转换、数据分析和模型开发，以及分析结果的展现这几个环节。可视化在每一个阶段都可以起到重要的作用。

1.3.1 数据获取与数据转换

数据必须能够通过获取、整合、转换成为适合进行处理的格式，这是任何分析的基础。用户需要分析的数据往往是以多种形式存在的。这些数据可能以文本文件形式存在，可能存储在关系型数据库系统当中，也可能存储在Hadoop文件系统中。可视化分析在这一阶段可以通过友好交互的图形化应用界面定义数据获取的机制和规则，生成数据抽取的代码，直接利用生成的代码或基于生成的代码将数据从各种不同的数据源当中高效地抽取出来。

数据的转换是指通过一定的步骤将数据转化成为能够提供更多信息的形式。一般来说，数据转换可以分为两类。一类是根据业务规则生成分析需要的新的数据，例如根据银行账户的余额和交易的发生额生成账户的每日余额和日均余额；另一类是根据分析的需要对现有数据进行技术上的转换，例如通过共线性分析将某些冗余变量删除，或对某些变量进行