本样章直接来自笔者原稿，因此在个别文字及排版上会与正式出版物有所差异。

关于本书的更多内容及下载请访问：www.StatStar.com

@文彤老师

第十五章　淘宝大卖家之营销数据分析

学习前建议阅读	第一章　数据分析方法论简介，了解三种数据分析方法论的异同； 第二章　数据分析方法体系简介，对统计方法体系做一基本了解。
案例导读	在本案例中，随着竞争的日益激烈，淘宝大卖家张三希望能够从头建立会员数据库，并利用这些数据改善其店铺经营状况。初期张三希望能够对会员促销的效果进行提升，并进一步深入分析存在重购行为的买家具有怎样的基本特征。 利用 IBM SPSS Statistics 的直销模块，分析师利用 RFM 模型进行了历史数据的分析，筛选出了应当优先考虑的促销名单；随后又进一步对存在重购行为的买家的基本特征进行了定位，该结果将被用于随后进一步改善营销活动的效果。
分析方法	RFM 模型； 分类树（作为直销模块的后台方法被调用）。
案例中用到的分析过程	转换：计算变量、重新编码； 数据：选择个案、排序个案、标识重复个案、分类汇总、合并文件； 描述统计：交叉表； 直销：RFM 分析、生成对产品作出响应的我的联系人的概要文件。
学习后建议阅读	第十六章　超市产品购买关联分析，体验数据挖掘方法体系在营销分析中的应用。

15.1　案例背景

15.1.1　卖家张三

今天我们案例的主人公是淘宝大卖家***，为了保护隐私，也为了方便表述，我们就叫他张三吧。

张三，护肤品/彩妆类卖家，几经打拼，信用积累到皇冠，但也累得半死。每日深陷于护肤品行业的红海鏖战之中。

客观的讲，张三同学也有过很多次成功的爆款商品，在业内也树立了一定的名望，但现在淘宝上的竞争越来越激烈，爆款可以带来销量，却带不来多少利润。而促销、聚划算之类的活动作来做去，最后却发现钱都被开平台的马老板给挣走了！而且不知为什么，张三总觉得自己对老客户的资源发掘不够充分，虽然现在也有很多的销售来自于回头客，但他私下里和其他几个大卖家一交流，却发现无论是自己店铺的回头率还是回购的客单价，似乎总是应当还有上升空间，可自己却又不知道究竟应当怎么做，唉，纠结呀。。。前辈们总是说吃一堑，长一智，可为什么都吃了很多堑了，可长出来的仍然只有智齿呢？

这不，张三最近认真策划的一轮会员大促销活动又没有达到原计划的效果，利润低于预期。这几乎成了压垮张三信心的最后一根稻草。好在老婆大人深明大义，在对其轻声抚慰之后一语道破迷津：不能再只知道卖傻力了，除了大投广告、做好装修、做好客服这些全银河系的淘宝卖家都知道的事情之外，是不是也应当上一点更有技术含量的东西，发掘发掘店铺销售数据库里面有没有可用的信息？这真是一语惊醒梦中人，经过和店员们的讨论，请教了淘宝小二，特别是在认真学习三个代表、刻苦抓好三讲教育、深刻揭批三聚氰胺，以及接受了老婆大人高瞻远瞩的最高指示之后，张三决定对自身淘宝店铺积累下来的数据进行分析，争取打一个漂亮的翻身仗。而在搜索了自身的资源之后，我作为张三多年未见的同学，很荣幸的开始走进了老同学的淘宝店铺营销活动策划之中。下面我们就来看看张三同学是怎样利用 SPSS 中的直销模块这一工具，对自身的店铺数据进行最基本的开发利用的吧。

由于是草根创业，张三的店铺几乎没有任何数据库方面的概念或者准备，当然也就更谈不上完备的数据仓库了，暂时能加以利用的只有交易表、历史买家数据库、历史商品数据库这些几乎处于原始状态的数据表单。当然，张三明白想通过数据分析得到的信息也很明确，首先是下面两点：

- ✧ 如果下次再做会员促销，那么究竟哪些是最有可能对促销信息作出反馈的会员？换言之，我究竟应当优先考虑对哪些会员进行促销？
- ✧ 和在本店铺无重购行为的买家相比，在本店铺存在重购行为的买家具有怎样的特征？

本案例中用到的数据为 2011 年第二季度张三店铺的所有原始销售数据，交易表见文件卖家张三_交易表.sav，其中包括买单号、买家 ID、商品 ID、购买时间、总价、运费、商品

案例精粹样章

数等字段；买家信息见文件卖家张三_买家表.sav，其中包括买家 ID、买家性别、买家年龄、买家省份、买家城市、买家信用等字段。

出于保护客户隐私的需要，本案例中涉及到买家/卖家隐私的数据均已删除，且相应的买家 ID 也已进行技术处理，因此相应的分析结果仅作为方法演示用。

11.1.2 分析思路/商业理解

本案例是一个非常典型的基于营销需求而来的数据分析案例，从数据分析的角度而言，其实问题并不复杂，但关键点在于所有的分析都需要紧密围绕着淘宝店铺的运营需求展开。针对以上三点需求而言：

- ✧ 究竟哪些是最有可能对促销信息作出反馈的会员？这是一个标准的从历史客户群定位可能“最有价值”的客户的分析需求，在营销方面有很多模型或者方法可以实现，但是在拥有明确的历史交易数据表的情况下，最为简单易懂而且使用的方法非 RFM 模型莫属；
- ✧ 在本店铺存在重购行为的买家具有怎样的特征？这一个分析需求如果从统计建模的角度来讲，则基本类似于对重购行为进行预测建模，并从中寻找重购行为的影响因素。
- ✧ 购买本店铺产品的买家大致可以被分为哪些类型？这从营销的角度实际上就是一个市场细分问题，而解决市场细分的方法中比较常用的则是聚类分析；

出于本案例明确的营销需求，我们这里将不再事先进行系统的数据理解和数据准备，而是完全从实战的角度出发，按照真实环境中的考虑进行操作，并在这一分析过程中逐步深化对数据的理解。

15.2 利用 RFM 模型定位促销名单

15.2.1 RFM 模型简介

RFM模型的定义

RFM模型是衡量客户价值和客户创利能力的重要工具和手段，他通过一个客户的近期购买行为、购买的总体频率以及花了多少钱三项指标来描述该客户的价值状况。其名称RFM来自于最近一次消费（Recently）、消费频率（Frequency）、消费金额（Monetary）三个英文单词首字母的缩写，是广泛应用于客户关系管理（CRM）的一种模型。而在众多的CRM分析模式中，RFM模型几乎是引用最广泛的一个。构建RFM的三个基本指标含义如下：

1. 最近一次消费（Recently）：指用户上一次购买的时间，理论上，上一次消费时间越

近的顾客应该是比较好的顾客，对提供即时的商品或是服务也最有可能会有反应，因此它对营销来说是一个重要指标，涉及吸引客户，保持客户，并赢得客户的忠诚度。这也就是为什么0至6个月的顾客收到营销人员的沟通信息多于31至36个月的顾客。当然，最近一次消费的过程是持续变动的。在顾客距上一次购买时间满一个月之后，在数据库里就成为最近一次消费为两个月的客户。反之，同一天，最近一次消费为3个月前的客户作了其下一次的购买，他就成为最近一次消费为一天前的顾客，也就有可能在很短的期间内就收到新的折价信息。

2. 消费频率（Frequency）：是顾客在一定时间段内的消费次数。最常购买的消费者，忠诚度也就最高，增加顾客购买的次数意味着从竞争对手处偷取市场占有率，由别人的手中赚取营业额。根据这个指标，分析者一般会把客户分成五等分，这个五等分分析相当于是一个“忠诚度的阶梯”（loyalty ladder），其诀窍在于让消费者一直顺着阶梯往上爬，把销售想像成是要将两次购买的顾客往上推成三次购买的顾客，把一次购买者变成两次的。

3. 消费金额（Monetary）：消费金额是对营销效果最直接的衡量指标，也可以验证“帕雷托法则”——公司80%的收入来自20%的顾客。

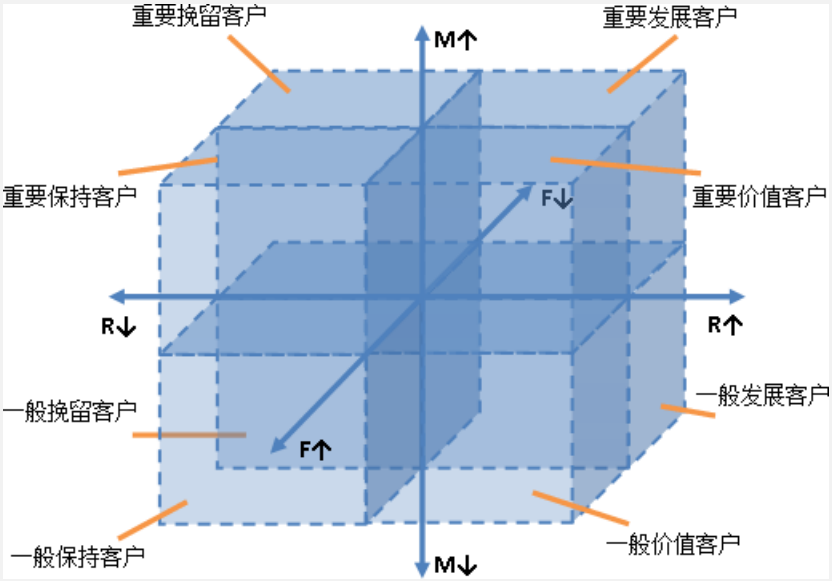

图15.1　基于RFM模型的客户细分框架

RFM的用途

RFM模型最大的价值在于可以从所有的历史客户群中迅速定位那些可能“最有价值”的客户，并通过随后及时的联络沟通，将其潜在购买转化为实际购买行为，从而进一步的增强客户忠诚度，封杀竞争对手的市场空间。对于大多数企业而言，在进行营销宣传时都会受到预算或者投入产出比的限制。例如营销活动的预算不多，只能提供服务信息给2000或3000个顾客，那么是将信息发给贡献40%收入的顾客，还是那些不到1%的顾客？这类情形就是RFM等模型的用武之地了。

RFM非常适用于生产多种商品的企业，而且这些商品单价相对不高，如消费品、化妆品、小家电、录像带店、超市等；它也适合在一个企业内只有少数耐久商品，但是该商品中有一部分属于消耗品，如复印机、打印机、汽车维修等消耗品；RFM对于加油站、旅行保险、运输、快递、快餐店、KTV、行动电话信用卡、证券公司等也很适合。

RFM分析原多用于传统营销、零售业等领域，只要任何有数据记录的消费都可以被用于分析。而对于电子商务网站而言，由于其网站数据库中记录的交易信息更加详细，因此同样可以运用RFM分析模型进行数据分析，并且相应的分析深度还可以基于数据的丰富程度做进一步的拓展，尤其对于那些已经建立起客户关系管理（CRM）系统的网站来说，其分析的结果将更具意义。

RFM模型的局限性

RFM虽然很有用，但也不可以用过头，从而造成高交易的客户不断收到信函。每一个企业应该设计一个客户接触频率规则，如购买三天或一周内应该发出一个感谢的电话或Email，并主动关心消费者是否有使用方面的问题，一个月后发出使用是否满意的询问，而三个月后则提供交叉销售的建议，并开始注意客户的流失可能性，不断地创造主动接触客户的机会。这样一来，客户再购买的机会也会大幅提高。

RFM分析也存在着明显的缺陷，从统计的角度讲，他其实根本算不上一个模型，只是在原先传统商业模式下数据比较匮乏的时候的一种简单快速的CRM方法而已。在当前数据仓库普遍建立的情况下，分析人员显然可以大大超越RFM的简单框架，真正去利用更丰富的数据信息，使用复杂的数据分析或数据挖掘模型来对高价值客户做到更为准确的定位。此外，RFM模型只能分析有交易行为的用户，而对访问过网站/商场但未消费的用户由于指标的限制无法进行分析，这样就无法发现潜在的客户。而对于电子商务网站而言，由于网站数据的丰富性——不仅拥有交易数据，而且可以收集到用户的浏览访问数据，可以扩展到更广阔的角度去观察用户，相信已经阅读过前面各章节的读者一定会对此非常清楚。

15.2.2 对数据进行 RFM 模型分析

下面我们就来对张三店铺第二季度的销售数据进行 RFM 分析，显然，建立 RFM 模型所需的交易时间、价格、重复频次均已包含在交易数据表中，因此首先打开交易数据表，然后就可以进行后续的操作了。

尝试初步构建 RFM 模型

这里我们首先按照默认的模型设定方式尝试建立 RFM 模型，并根据分析结果来确定应当如何调整模型选项，操作如下：

1. 直销➔选择方法；

2. 选择左上角的“帮助标识我的最佳联系人（RFM 分析）”，继续；
3. RFM 分析：数据格式：选择数据格式为交易格式；
4. 将购买时间选入交易日期框，总价选入交易金额框，买家 ID 选入客户标识符框；
5. 确定。

图 15.2　直销模块的选择方法对话框

在直销模块的选择方法对话框中，我们可以看到其中的方法完全是按照营销需求的种类在进行排列，而具体背后所采用的统计模型则并未作为重点加以展示，这正是体现了这一模块在设计上的营销需求导向。

图 15.3　RFM 分析的主对话框

在默认设定下，RFM 模型的输出结果其实非常简单，主要就是下面的这两张图型，以及计算出的 RFM 数据表。首先来看 RFM 块计数图，该图形显示的是按照设定的离散化方法所生成的块分布。而每个直条的高度则代表所生成的合并 RFM 得分组的案例数。模型默认会分别对购买频次、时间、金额的大小将案例尽量五等分，并依次给出 1-5 分。这样三个维度交叉，就会将所有样本拆分入 5×5×5=125 组。而分析者随后的工作则是从这 125 组中挑选出最有价值的若干组来实施相应的营销操作。

虽然从原理上 RFM 方法会尽量将样本进行五等分，但是在实际分析中，由于经常会出现大量相等数值的情形（特别是购买频次这一指标），这会导致各组的样本量分配出现偏差。以本案例为例，可以看到频率指标实际上并未被分为五组，而是被划分为得分为 3、5 的两组，并且从直条高度即可看出，3 分组的频数要远高于 5 分组。这样悬殊的样本组间分配显然会影响 RFM 的分组效果，提示我们应当对模型的默认设定进行修改已获得更加均匀的分组结果。

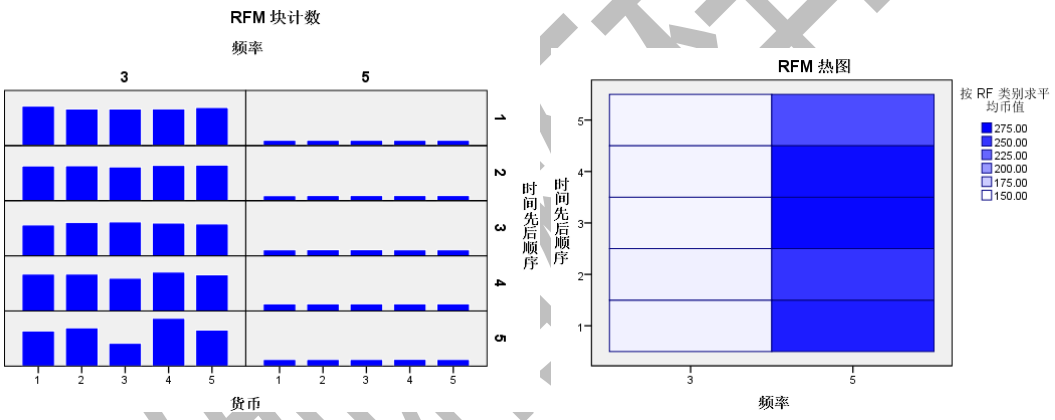

图 15.4　RFM 块计数图和 RFM 热图

除块计数图外，SPSS 还会给出 RFM 热图，用于表示不同购买时间×购买频次分组时该组段的平均购买金额分布，以协助使用者迅速确定高价值组段所在的区域。图中颜色越深的区域表示平均货币值越高，换言之，时间先后顺序和频率得分在深色区域中的客户，比时间先后顺序和频率得分在浅色区域中的客户的平均购买金额要高。从图中可以看出，频率分值为 5 的几个组段其平均购买金额均明显高于频率分值为 3 的组段，这显示出在本数据中，购买频次与购买总金额的数量关联要明显高于时间顺序。

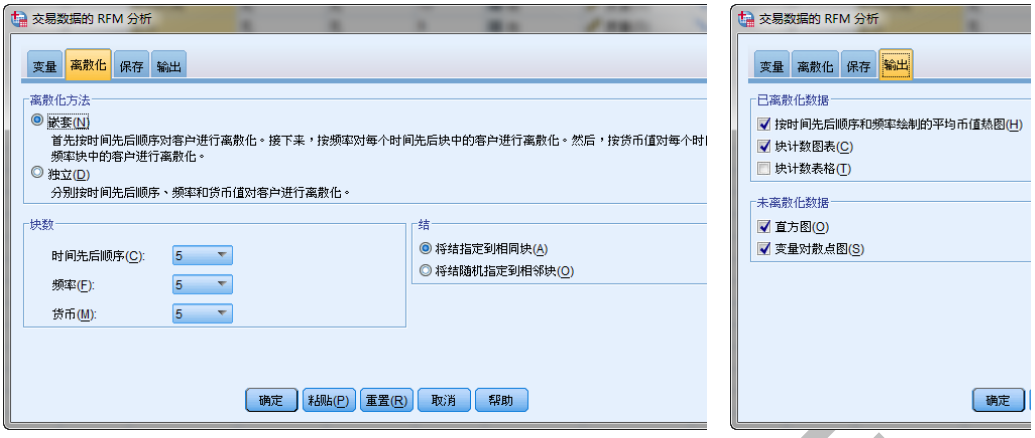

图 15.5　RFM 模型的离散化选项卡和输出选项卡

调整 RFM 模型的离散化选项

在得到上述 RFM 模型的初步分组结果之后，我们可以基于该结果直接筛选出评分较高的组进行营销操作，例如在本例中，显然可以将购买频率得到 5 分的各组都纳入营销操作范围。但是这样不均衡的分组结果显然会影响 RFM 模型的应用效果，因此需要进一步寻求对案例能够进行更为均匀的细分的分组方式，在本例中，相应更改默认选项的操作如下：

1. 离散化选项卡：结框组：将结指定到随机相邻块；
2. 确定。

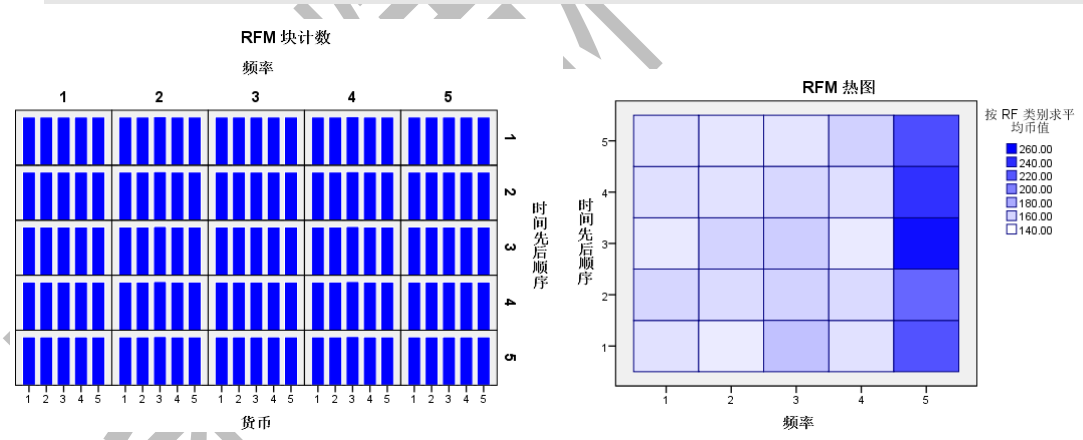

图 15.6　应用离散化选项之后的 RFM 块计数图和 RFM 热图

通过将结指定到随机相邻块，就可以有效解决分组不均匀的问题。现在从 RFM 块计数图的直条高度即可看出，所生成的 125 个组段其样本量基本上是相同的。而 RFM 热图则进一步明确提示其中频率得分为 5 分（注意这里的 5 分组段不等同于前面分析中的 5 分组段）的几个组段其平均购买金额明显更高一些，而其中又以时间顺序得分为 3-5 分的三个组段更高一些。

进一步展示时间、频次、金额间的数量关联

通过上面的分析，我们已经得到了 RFM 模型的分组结果，但是为了能够更充分的理解数据的含义，还是应当对时间、频次、金额三个指标间的数量关联进行了解，这可以通过模型的相应选项来实现，操作如下：

1. 输出选项卡：未离散化数据框组：选择直方图和变量对散点图；
2. 确定。

根据上面的设定，系统会在结果中输出上述三个变量的直方图和两两散点图。直方图显示用于计算时间先后顺序、频率和金额这三个字段的值的相对分布。每个直方图的水平轴始终采用左侧为较小值、右侧为较大值的顺序。但对于时间先后顺序，图表的解释依赖于时间先后顺序测量的类型：日期或时间间隔。对于日期，左侧条代表更“早”的值（即较远日期比较近日期的值更小）。对于时间间隔，左侧条代表更“近”的值（即时间间隔越小，交易离现在越近）。在本例中可以看到时间的直方图呈负偏态分布，由于本例中时间为日期格式，因此这意味着第二季度中 6 月份的交易数量是要高于 4 月份的。至于频率和金额字段，二者均呈强烈的正偏态分布，说明大部分的买家都属于频率较低、金额较低的类型，这显然非常符合实际情况。

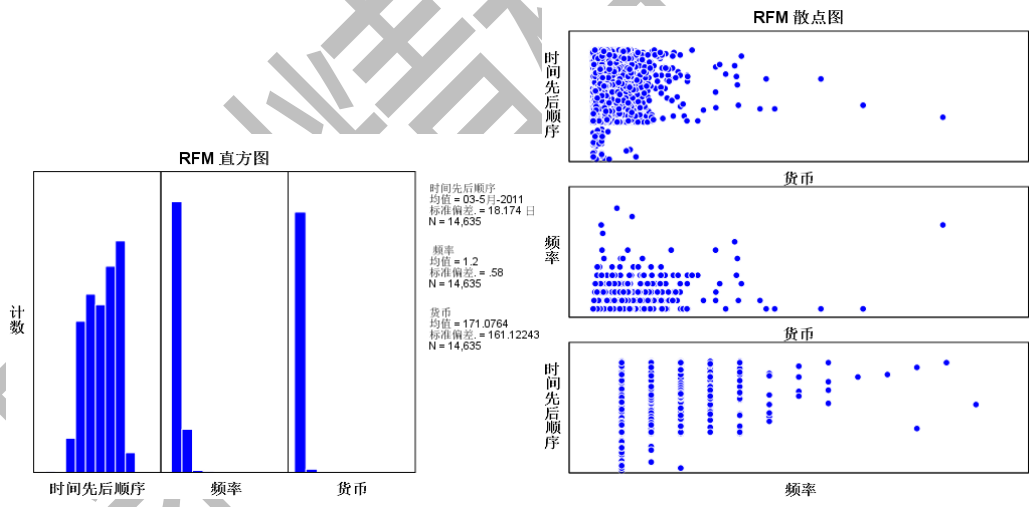

图 15.7　时间、金额、频率的直方图和散点图

除直方图外，两两散点图则显示出时间先后顺序、频率和金额这三个变量之间的关系。和直方图的情况类似，时间先后顺序轴的解释也依赖于时间先后顺序测量的类型：日期或时间间隔。对于日期，越接近原点的点代表离现在越远的过去日期。对于时间间隔，越接近原点的点代表越“近”的值。在本例中可以看出金额较高的交易似乎更多的出现在第二季度早期而不是后期；与此同时，高频率购物者每次的购买金额实际上是相对偏低的；至于第三幅散点图显示出的累计购买频率较高的购物行为其最近交易日期更多出现于第二季度后期的现象则属于合乎逻辑的数据错觉，原因请读者自行思索。

得到 RFM 数据表并加以应用

上述 RFM 模型在计算完毕后，会生成一个新的活动数据集，其中记录了每个买家 ID 所对应的 R、F、M 分组得分，以及最终总的 RFM 得分。请注意，由于在大量的研究中发现这三个变量在决定消费者购买预期时的重要性依次是时间、频率、金额，因此总 RFM 得分的计算方式为：R*100+F*10+M，相应的 125 个组段则分别得到 111～555 分，研究者随后需要做的工作就是按照自身需求筛选出 RFM 总分高于设定界值的买家 ID 以进行后续的营销操作即可。例如在本例中，就可以将 RFM 总分高于 500 分的买家 ID 筛选出来，具体操作为：

```
SELECT IF RFM_得分>500.
EXECUTE.
```

注意按照上述程序，数据集中 RFM 得分低于 500 分的案例将会被全部删除，如果只是希望在分析中滤掉这些案例而非删除，则应当使用 filter 语句，或者在数据➔选择个案对话框中进行相应的操作。

	buyer_id	最_近_日期	交易_计数	金额	崭新_得分	频率_得分	消费金额_得分	RFM_得分
1	10007528	07-May-2011	1	198.00	3	3	4	334
2	10010926	13-Apr-2011	1	188.00	1	3	4	134
3	100324521	23-May-2011	1	118.00	5	3	2	532
4	100386022	30-Apr-2011	1	42.00	3	3	1	331
5	10042051	16-Apr-2011	1	108.00	2	3	2	232
6	100432931	24-May-2011	1	236.00	5	3	5	535
7	100437524	27-May-2011	1	236.00	5	3	5	535
8	10045585	08-Apr-2011	1	238.00	1	3	5	135
9	10054602	06-May-2011	2	118.86	3	5	2	352
10	100570823	12-Apr-2011	1	324.00	1	3	5	135

图 15.8　生成的 RFM 数据表

15.3　寻找有重购行为买家的特征

（本样章略去本节）

15.3.1　数据理解与数据准备

对买家表中的变量进行清理和转换

对交易表中的变量进行清理和转换

合并买家表与 RFM 数据表

15.3.2　利用直销模块寻找重购人群的特征

15.4　总结与讨论

15.4.1 可使用的其他营销分析方法

现在卖家张三已经回答了至少两个在具体业务中比较关心的营销问题，基于控制篇幅的考虑，这里我们就不再继续展开讨论下一步的营销分析了。但是很显然，随着营销活动的继续开展，必然会有更多的分析需求出现。简单地说，在淘宝卖家的业务流程中至少还可以研究如下几方面的问题：

- ✧ 客户细分：基于买家的历史购买记录，根据不同买家的购买特征对其进行细分，从中发现店铺的客户主要可以被分为几种类型，并据此进一步制定差异化的营销活动。该分析可以在直销模块的“将我的联系人分段到群集”中完成。
- ✧ 潜在买家定位：本次活动中促销名单是采用 RFM 模型进行筛选，随着历史数据的积累和完善，后续的营销活动完全可以换用更精确一些的方法来确定促销名单，即基于历史营销活动的数据来建立针对不同种类促销活动高响应概率人群的预测模型，并将该模型应用于未来促销活动名单的筛选分析之中。当不存在同类历史营销数据时，在时间许可的情况下，也可以先抽取一小组买家进行营销测试，然后基于测试结果数据来建立相应的预测模型，并将其应用于后续的正式营销活动。该分析可以在直销模块的“选择最有可能购买的联系人”中完成。
- ✧ 营销活动的效果比较与改善：针对同一个营销目标，卖家往往可以有多种选择，例如是打 9 折还是立减 10 元，是一视同仁还是只针对三星级以上买家促销等。为更准确的回答此类问题，完全可以将买家随机分为若干组，每组采用某种确定的营销手段来进行推广，然后比较各组的营销效果，从而得出更加客观的结论。在时间许可的情况下，这一操作可以作为营销测试手段。如果时间条件不具备，也完全可以直接应用到正式营销活动中去，然后将分析结果用于未来营销活动的改善上去。该分析可以在直销模块的“比较活动效果（控制包装检验）”中完成。

SPSS 的直销模块还可以帮助卖家完成更多的营销分析和操作，这里不再一一列举，有兴趣的读者可以自行查阅该模块的帮助文档了解更多细节。

15.4.2 研究总结

明眼人都可以看出，本案例中所采用的分析方法，或者说直销模块中可能会采用的各种方法其复杂程度相差极大。RFM 分析、聚类分析在很多统计学者眼中几乎就不能算是统计方法，学过算数的小学生恐怕都能鼓捣出结果来。而 logistic 回归之类的方法又是比较复杂高深的统计模型，系统学习过统计学的人也未必能够正确使用和解读。但为什么这些难易程度相差极大的方法能够被放置在同一个直销模块中呢？

案例精粹样章

实际上，本案例的分析需求应当被归类为比较典型的数据挖掘分析需求，也就是完全以实际的商业需求为导向，并以分析结论能否满足商业需求作为项目成功的唯一评判标准。这一样来，无论方法学的简单或者复杂，只要能够满足分析需求就可以考虑加以使用。实际的营销活动中情况千变万化，对响应时间的要求也可能很高，那种慢工出细活的复杂建模思路不但可能无法满足时间要求，也不见得能够得到更好的结果。这也是为什么在营销分析中，在方法学上如此简单的 RFM 分析可以大行其道的原因－－这东西简单易懂效果又好，干嘛不用？还是那句话：无论白猫还是黑猫，只要抓住老鼠，就是好 Hello Kitty。

案例精粹样章