互动出版网 www.china-pub.com 专业图书网上第一专营店 www.china-pub.com 独家提供样章

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）。

互动出版网 www.china-pub.com 专业图书网上第一专营店 www.china-pub.com 独家提供样章

	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。

互动出版网 www.china-pub.com 专业图书网上第一专营店 www.china-pub.com 独家提供样章

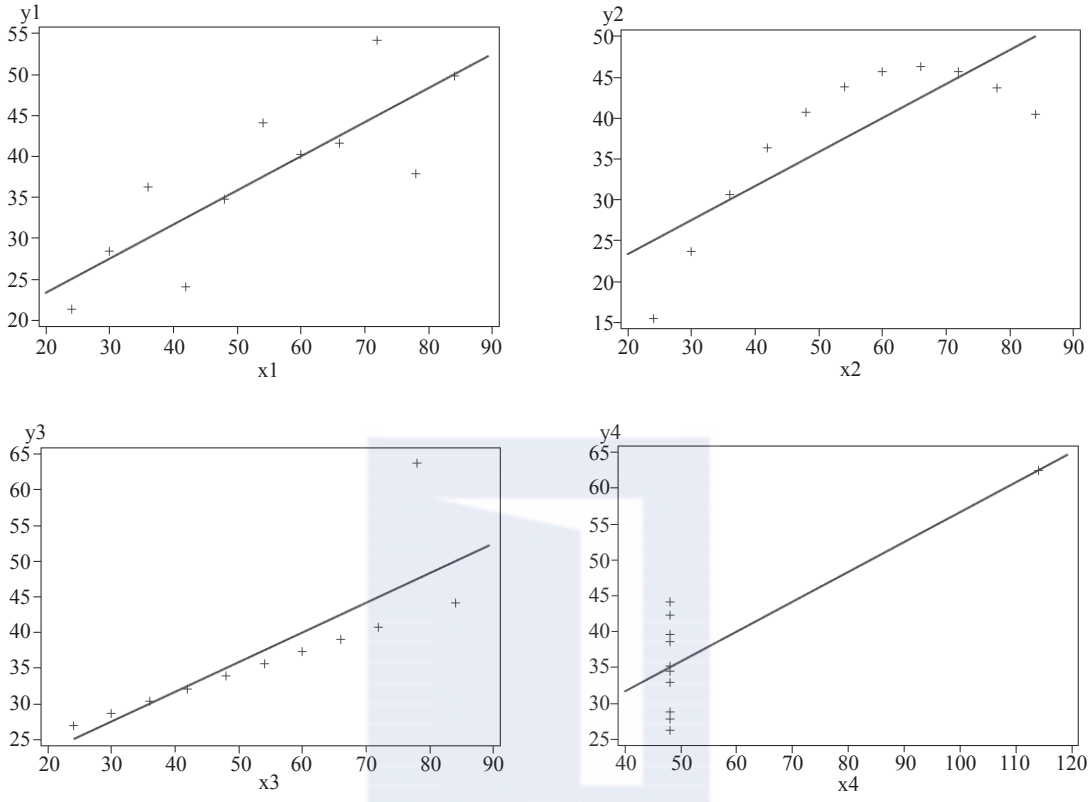

图 1-3　变量关系可视化展现

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

互动出版网 www.china-pub.com 专业图书网上第一专营店 www.china-pub.com 独家提供样章

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

互动出版网 www.china-pub.com 专业图书网上第一专营店 www.china-pub.com 独家提供样章

Log 变换。数据转换的过程同样可以用可视化的方式辅助实现。

1.3.2　高级分析与模型开发

可视化分析技术同样可以为高级分析提供可视化的交互界面和用以分析的辅助图像。在这个阶段，可视化分析技术可以提供诸如散点图、盒须图、热力图、气泡图等对分析极有帮助的各种图像。可视化还可以使分析人员借助友好的交互界面使用高级分析技术包括机器学习技术来建立模型并进行模型评估等工作。总之，可视化使得用户不需要大量编码就可以使用各类高级分析技术，从而使得分析的门槛大大降低，普通的业务或技术人员经过一定的培训也可以进行高级分析。

1.3.3　分析结果展现与模型应用

只有分析的结果能够及时有效地和决策人员共享，这样的分析才是有意义的。可视化使得分析过程、结果可以被生动灵活地展现和分享，从而帮助决策者理解分析结果。如果分析的结果是一个模型，那么这个模型就应该能够方便快捷地部署并应用起来，只有这样模型才可以在决策过程中发挥作用。可视化技术能够简化模型部署的过程，并且使得监控模型的工作变得更简单。

人们谈及数据可视化分析，一个常见的误区是认为可视化分析就是报表和图形。实际上数据可视化分析涵盖了数据分析的各个层面，具体而言，可以分为下面不同的 8 个层次：

- 标准报表：标准报表是一个组织机构甚至一个行业所需要的基本固定报表或数据图表，可以回答诸如“发生了什么”以及“什么时候发生”这样的简单业务问题。
- 即席报表：即席报表可以允许使用人员在一定的范围内输入条件信息然后即时地按照输入的条件得出报表或图形报告。可以对于“在哪里发生”“发生频率”以及“多少”这样的业务问题进行回答。比如业务人员会希望立刻看到对于某一个区域在某个时间段的销售情况。
- 钻取查询：钻取查询的功能就是联机分析处理（OLAP）所提供的功能。它可以让业务人员从不同的业务维度分析结果，并对数据进行钻取从而分析问题发生的根本原因。
- 警报：警报信息可以在发生特定问题的情况下通知相关人员。比如，当销售目标低于预期时，销售管理人员会收到提醒，提醒的方式可以通过电子邮件，也可以通过仪表盘等方式。

这四个层次的分析基本上就是传统意义上的报表，这些分析可以根据数据对于已经发生的事情生成报表。但是这些分析的不足也是显而易见的，即它们都不能提供关于未来的任何分析。如果业务部门需要了解更为复杂的关于未来的预测性的分析，那么就需要依赖高级分析，即下面的四个层次的高级分析。

- 统计分析：基于数据，统计分析可以应用一些较为复杂的分析模型，如回归。业务

互动出版网 www.china-pub.com 专业图书网上第一专营店 www.china-pub.com 独家提供样章

人员通过回归模型可以了解为什么某件事情会发生，以及影响该事件发生的各个因素所占的权重。

- 时间序列预测：时间序列预测可以用来分析按时间顺序生成的数据列。例如可以帮助零售商预计某个商品未来一段时间在各个门店的需求。这种需求预测可以帮助零售商以适当的成本提供定量的商品以应对客户对于不同产品的需求。
- 预测模型：假设某个公司拥有上千万的客户，如果该公司希望展开一次市场营销活动，那么哪些客户会对该营销活动积极响应？或者该公司希望了解自己的客户中有哪些客户可能流失？预测模型正是用来回答类似这样的问题的。预测模型已经成功应用于许多领域，比如风险评估、欺诈监控和数据库营销等。
- 优化：优化可以解决在资源约束的情况下，如何得到最佳产出的问题。比如在市场营销活动中，业务人员往往面临营销经费固定，营销人员数量有限的约束，如何在这样的约束情况下取得最佳的市场营销效果，市场营销优化就可以给出最优的营销策略。另外在供应链领域，库存如何优化也是一个常见的问题，库存过高会给整个供应链带来资金的压力，库存过低则可能不能及时满足客户的需求，库存优化则可以在满足客户需求的情况下将整个供应链的库存尽可能降低。

上面这四类高等分析可以基于数据进行统计分析、预测和优化，因此可以提供预测性的洞察力，从而回答更为复杂的业务问题。为了应对日益复杂的业务问题，数据可视化分析软件需要提供从标准报表到预测以及优化的全方位的解决方案。

1.4 可视化分析面临的挑战与应对

1.4.1 可视化分析面临的挑战

在大数据时代，数据可视化分析面临巨大的挑战，企业需要更新的数据分析平台以满足不断发展的数据分析需求。新的数据分析平台要能够对海量的数据进行快速的处理、探索和高级分析。在大数据时代，企业的数据量快速的增长，企业所得到的数据会呈现多样性，企业需要处理的业务问题也越发复杂，因此企业需要拥有一个统一的数据分析平台去分析大量的各种不同类型的数据，并且能够快速地对任何分析问题进行处理。新一代的数据分析平台还应具有扩展性，能够随着数据量和分析人员的不断增加，而提供可扩展的数据分析能力。

数据可视化分析面临的另外一个挑战就是对数据和分析过程的全面管控。各类开源技术的涌现使得人们可以任意搭配各类不同的开源技术平台对数据进行可视化分析，大大推动了可视化分析的普及。可是各类开源技术经常不能够很好地集成，因而导致对于数据和分析过程缺乏有效的管控，这会随之而来增加各种风险，比如维护成本、管控成本和合规风险等。在企业内部对于数据进行处理的工具经常是多样的，这会使得跟踪数据的来源变

互动出版网 www.china-pub.com 专业图书网上第一专营店 www.china-pub.com 独家提供样章

得越发复杂，此外人们还需要知道现在正在使用的模型是哪一个版本、是否是最新发布的模型以及是否能够定期对模型进行更新。所有这些都表明一个能够对所有数据和整个分析流程进行管控的数据可视化分析平台是十分必要的。

数据可视化分析要能够解决整个数据分析生命周期中的各种问题。完整的数据分析过程包括数据准备、数据探索、数据变换和变量选择、建立模型、验证模型、部署模型及持续评估和监控模型表现。任何可视化分析的基础都是高质量的数据，在数据准备阶段，数据可视化分析要求能够获取各种不同类型的数据。客户的数据可能以多种形式存储，可以是传统的 Excel 文件、文本文件、关系型数据库、电子邮件、各类应用系统、网页、社交媒体流，也可以是 Hadoop、Cassandra 等分布式存储系统，数据可视化分析要能够帮助用户通过可视化的界面获取各种不同类型的数据，并且能够对数据进行有效的整合。

在数据探索阶段，需要有可视化的界面帮助各种水平的用户对数据进行探索。用户可以借助可视化的界面和各种不同类型的可视化图形对数据进行探索性分析，生成各种直观的报表、图形。在建立模型阶段，对于不擅长编程的用户，数据可视化分析平台也应该提供交互式的界面帮助用户建立各种预测模型。借助可视化的界面和可视化的报表、图形，拥有不同知识水平的人员都可以充分利用数据分析的能力，得到并且分享数据分析的结果。

数据可视化分析平台还应该提供全面的模型存储、模型监控、模型执行能力。数据可视化分析平台应该对各种类型的模型提供统一的模型管理界面，用户只需要一次性导入模型，然后就可以在各种不同系统当中使用该模型。模型管理界面还应当具备持续的模型监控能力，当模型的表现开始出现明显衰减时，建模人员可以得到提醒进而重新训练并且寻找冠军模型。可视化数据分析平台还应该提供完整的模型执行的能力，各种模型能够轻易地部署到各个生产环境中。

1.4.2　SAS 的可视化分析实现

为了应对大数据时代的可视化分析挑战，SAS 公司推出了新一代的高性能内存分析平台。该平台的架构易于在公有云、私有云以及其他操作系统安装部署，因此具有良好的可扩展性。同时，该平台提供了基于内存的、分布式的处理能力，可以让多个用户同时对大量数据进行处理，解决复杂的分析问题。该平台为拥有 SAS 技能以及其他语言技能的人员提供了一个开放统一的平台，拥有不同编程语言技能的人都可以借助该平台解决各种复杂的分析问题。

SAS 推出的统一的数据可视化分析平台具有以下特点，能够很好地解决大数据时代所面临的数据分析挑战。

- 全面的分析管控。SAS 新一代可视化数据分析平台为企业级的数据分析提供了必要的管控。它可以让企业对独立分散的数据分析流程进行有效的管理，可以帮助企业内部的数据分析人员管理各种不同语言生成的模型，同时帮助 IT 部门对于所有的分析过程进行监管。它还可以对模型版本、模型权限，数据源等信息进行统一的管理，

互动出版网 www.china-pub.com 专业图书网上第一专营店 www.china-pub.com 独家提供样章

从而确保企业在整个分析流程中所使用的数据、模型和结果都是可以信赖的。

- 可以信赖的分析结果。可视化数据分析的结果会指导商业决策，在风险、欺诈和网络安全等领域，数据分析结果的精准性至关重要。从简单的线性回归到复杂的机器学习算法，SAS 提供了广泛的经过各个领域实际验证的分析功能，这些分析功能经过严格的测试，在 SAS 的不同版本之间的运行结果保持一致。
- 可视化分析界面。SAS 新一代的可视化分析平台提供了友好的可视化分析界面。普通用户可以通过可视化的界面进行数据准备、数据探索以及模型建立，不需要了解编程语言就可以使用 SAS 强大的数据管理和数据分析能力。而具有编程能力的用户可以使用自己所习惯的语言进行编程，SAS 新一代的可视化分析平台支持用户通过 Python、Java、R 或者 Lua 语言去调用 SAS 强大的数据管理和数据分析能力。
- 人人可以使用数据分析。SAS 新一代的可视化分析平台所提供的自助式的数据准备、数据探索、模型建立等功能可以让企业内部的非技术人员都能够使用 SAS 提供的强大分析能力，将数据转化成为可以信赖的决策。
- 模型部署简单化。SAS 新一代的可视化分析平台提供了对于各种语言的模型进行存储、执行和监控的统一平台。企业可以轻易地部署模型，然后在企业内部的不同生产系统中调用该模型。
- 高性能。SAS 新一代可视化数据分析平台所采用的分布式的基于内存的架构使得数据处理的速度大大加快，以往需要几个小时运行的工作采用新的数据分析平台后往往几分钟就可以得到结果。用户在使用 SAS 函数的时候也无需将数据从 Hadoop 中进行抽取，SAS 函数支持在 Hadoop 内部运行。

借助于 SAS 的新一代的可视化分析平台所拥有的技术优势，SAS 采用不同的产品和技术去满足上面提到的数据分析的 8 个不同层面的需求。

- SAS 可视化分析（SAS Visual Analytics）。针对一般的报表和钻取查询，SAS 提供了可视化分析产品。SAS 可视化分析借助 SAS 基于内存的分析引擎，支持从不同的数据源将数据加载到内存当中，快速检索海量的数据，并且最终以报表的形式展现。SAS 可视化分析分为三个模块：SAS 可视化数据生成器（SAS Visual Data Builder）、SAS 可视化探索器（SAS Visual Analytics Explorer）、SAS 可视化设计器（SAS Visual Analytics Designer）。SAS Visual Data Builder 通过可视化的界面为业务人员提供了访问不同数据源的能力，用户可以访问数据库当中的表、文本文件、存储在 Hadoop 当中的数据，并且将这些数据加载到内存分析引擎中。SAS Visual Analytics Explorer 允许用户对加载到内存中的数据进行交互式探索，并且可以生成各种不同的图形和表格。SAS Visual Analytics Designer 可使用户轻松创建各种不同类型的报表和仪表盘，这些报表支持过滤和高亮这样的交互式操作。
- SAS 可视化统计（SAS Visual Statistics）。SAS Visual Statistics 提供交互式的界面，用户通过界面可以快速建立预测模型。SAS Visual Statistics 充分利用了 SAS 的基于

互动出版网 www.china-pub.com 专业图书网上第一专营店 www.china-pub.com 独家提供样章

内存的分析引擎，可以快速地对大量数据进行分析，允许用户在短时间内对多个模型进行开发和验证。用户可以方便地对模型进行评估，将选定的冠军模型投入到生产环境中，最终让分析模型落地的时间大大缩短。SAS Visual Statistics 针对预测模型可以进行线性回归模型、逻辑回归模型、广义线性模型和决策树模型。此外 SAS Visual Statistics 还提供了聚类模型。

- SAS 可视化数据挖掘和机器学习（SAS Visual Data Mining and Machine Learning，简称 SAS VDMML）。SAS VDMML 给用户提供了数据挖掘和机器学习的工具。它集成了获取数据、数据转换、特征工程、探索性数据分析、建立模型、比较模型和生成评分代码等所有数据挖掘和机器学习所需要的功能。在这单一平台上，用户可以针对监督学习和非监督学习使用统计学方法、机器学习算法以及文本分析算法。它提供的交互式的界面让普通业务人员可以同样使用 SAS 强大的高级分析功能。

此外 SAS 还推出了基于新一代数据分析平台的通用型解决方案——SAS 可视化调查（SAS Visual Investigator）。它可以使信息分析人员和调查员减少误报，简化调查过程，打击欺诈行为并改善客户细分。SAS Visual Investigator 支持将不同类型、大小和位置的数据集中起来，实现数据搜索、查询。它还可以利用高级分析方法对事件进行风险评估，帮助调查人员将精力集中在高风险的事件上，并且支持将实体间关系进行网络可视化从而发现有价值的隐藏信息。SAS Visual Investigator 可以广泛应用在各个领域：欺诈探测、风险分析、零售损失预防、机器性能监控。

1.5　本章小结

本章介绍了数据可视化分析的含义、意义以及近些年逐渐兴起的背景，同时明确了数据可视化分析可以在数据分析的每个阶段发挥积极的作用，并且说明了数据可视化分析不仅仅只是报表，还应该包含高级分析。本章最后探讨了可视化分析所面临的各类挑战以及 SAS 如何应对这些挑战。