第 5 章

探 索 数 据

如果你无法为数据科学家提供可靠的数据……那么你注定会得到一堆糟糕的分析。[①]

If you tell a data scientist to go on a fishing expedition... then you deserve what you get, which is a bad analysis.

——托马斯·C.里德曼(Thomas C.Redman),
“数据医生”,《哈佛商业评论》撰稿人

① 埃米·加洛主编的《哈佛商业评论管理必读:数据分析》第 10 章中引用了这句话。

在会议室中汇报和展示的数据项目总是显得非常简单明了，但实际上却并非如此。股东们看到的往往只是一个精心准备的幻灯片，按照从问题到数据再到结论的顺序依次呈现。但在这当中，有些想法被忽略了，包括数据团队在得到结果的过程中做出的那些重要决策。一个优秀的数据团队从不会只向一个方向前进，而是会根据从数据中获得的新发现灵活地改变方向。随着项目的推进，他们也会时不时地回到早期的想法，探索其他方向的可能性。

这类对数据反复检验探索的过程称为**探索性数据分析**（Exploratory Data Analysis，EDA）。这个概念在 1970 年由约翰·图基提出，指的是在采用复杂方法之前，先通过简单的汇总统计和可视化方法获得对数据的初步认知。[①] 图基认为，EDA 的过程就像侦探的工作流程一样，通过从数据当中获取更多的线索，确定下一步的行动。事实上，EDA 是另外一种“质询”数据的方式，它是任何数据工作当中必不可少的一部分，帮助人们确立并不断调整项目的方向。

5.1 探索性数据分析

有些人或许会对探索性数据分析的概念感到抗拒，因为它

① 参见约翰·图基《探索性数据分析》。

揭露了数据工作背后的实质(或艺术)。当两个数据团队拿到同样的数据和问题时,他们或许会沿着完全不同的路径探索,而最终得出的结论可能相同,也可能不同。这是因为,在探索的过程当中要做出太多的决策,不可能始终保持一致。每个人都会根据他们不同的背景、想法和工具,决定怎样才能最好地解决问题。

因此,本章将探索性数据分析视为一个持续的过程,每位数据达人都应该参与其中——不论你是第一手的数据工作者,还是在会议室当中听报告的商业领导。你将学会在探索数据的过程中该提出哪些问题,以及该注意哪些方面。

你是管理者还是负责人?

如果你是利益相关者、管理者,或相关领域的专家,那么请尽可能地与数据团队保持沟通,坦诚地对话,并做好进行多轮交流的准备。与他们一起设立正确的假设,不要让数据团队在不清楚商业背景的情况下就开始搜寻数据。否则,他们将会采取在统计学上,而非实际应用上更有意义的方式开展工作,而一个错误的假设将会危及后续的全部结论。

我们完全理解管理者不能像数据工作者那样密切地参与项目,但这也不意味着这样的情况不会得到丝毫改善。如果你是管理者,你不必事无巨细地参与,但也不能不闻不问。[①]

① 利益相关者也不应该事无巨细地参与,商业团队和数据团队之间应该保持信任。

5.2 培养探索心态

数据团队可以借助各类软件，通过汇总统计或可视化的方法，快速又廉价地对数据进行初步探索。但探索性数据分析并不应被视为一个工具清单。它更像是一种心态，应当贯穿数据工作的每一个阶段，即使你没有相关背景也可以参与其中。

引导性问题

接下来，我们将通过一个简单的情境来帮助你培养探索心态。例子中使用了一个常用于教学的数据集：埃姆斯住房数据（Ames Housing Data）。[①] 在这个例子中，我们将会简要展示探索性数据分析的过程。

尽管正确的数据探索路径并不单一，但你可以通过提出下面 3 个问题，引导你的团队得出有意义的结论。

- 数据是否能解答问题？
- 你是否能从数据中发现某些相关性？
- 你是否在数据中发现了新的机会？

让我们进入情境，依次研究这 3 个问题，讨论它们的价值，并分享你可能会遇到的困难。

背景设定

你在一家房地产行业的初创公司工作，需要为公司吸引顾客，但你发现很难与那些房地产行业的科技巨头竞争。比如说，

① 该数据集可以在 www.kaggle.com/c/house-prices-advanced-regression-techniques 网站下载。

美国的 Zillow 公司开发了著名的 Zestimate 软件，用来估算房屋的价值。这不仅为 Zillow 公司招揽了大量顾客，也为他们赢得了丰厚的利润。为了与之竞争，你的公司也要开发自己的预测软件。所以，你的任务是构建一个**模型**，以房屋的信息作为**输入**，房屋的估价作为**输出**。领导给了你一个数据集作为起点，这个数据集共有 80 多个变量，描述了 2006 年到 2011 年间美国艾奥瓦州埃姆斯市售出的几百座房屋的信息。

一时间接收如此大的信息量或许会使人吃不消，但之前提到的那些问题可以帮助你缩小范围，找到一个出发点。

接下来，让我们依次讨论这些问题。

5.3 数据是否能解答问题？

尽管我们可能很想将所有的数据直接扔到某个时下流行的算法当中去（例如第 12 章将会涉及的深度学习），但还是必须首先询问："这些数据是否能够解答问题？"很多时候，只需要简单地浏览数据，就可以找到答案。

建立预期，调动常识

假如让你说出估计房屋价格时需要使用哪些信息，相信你能说个八九不离十：房屋面积、卧室个数、卫生间个数、修建年代等。房屋买主往往会通过这些信息在你将要搭建的网站或平台上进行搜索。如果抛开这些信息去预测房价，将是空中楼阁。

当你打开展示数据的文件后，可以看到变量名和变量类型。那里有我们预期之中的变量，外加一些对预估房价有帮助的定序变量（房屋整体质量，从 1 到 10，10 代表"极佳"）、定类变量（地段类型），以及一些其他变量。这个数据集看上去比较正常。

接下来，你或许会想到检查变量的取值。它们覆盖了你想要分析的情境吗？比如说，如果你看到一个叫“建筑类型”的变量，发现其中只包括独栋住宅，却没有公寓或联排式住宅，那么相较于 Zillow 公司，你的模型就更加局限。Zestimate 可以为公寓估价，而如果你的数据当中没有公寓，那么你的公司就无法在这方面提供可靠的估值。

这里可以得到的教益是：避免本章开头的引言中提到的问题，确认我们得到的数据足够可靠，能够回答当前的问题。

变量取值是否符合直觉？

你可以借助软件生成一系列汇总统计，再结合问题的背景进一步分析数据，检查汇总统计的结果是否符合你对当前问题的直觉理解。可视化是探索性数据分析的重要组成部分，我们可以借助它来发觉数据中的异常之处。

数据可视化示例

接下来，我们会给出一些探索性数据分析的例子，其中用到了直方图、箱形图、条形图、折线图和散点图。如果你熟悉这些图表提供的信息，可以跳过这部分内容。

直方图可以显示连续数值数据是如何分布的。图 5.1 是房屋售价的直方图，可以看到，有 125 户的售价在 200 000 美元范围内，而右侧的长尾显示了那些最贵的房屋。右侧的长尾使得房屋均价(181 000 美元)高于售价中位数(163 000 美元)。当存在少数非常昂贵的房屋时，售价平均数就会高于中位数。

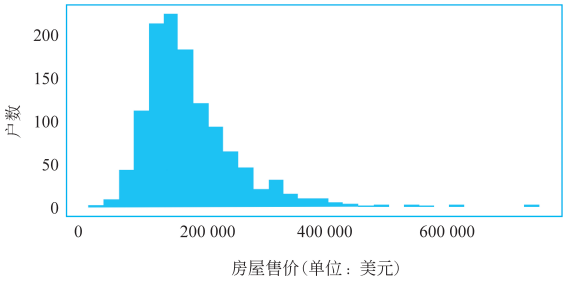

图 5.1 房屋售价直方图

直方图能帮助我们发现异常，如果我们看到售价为负值(买房反而赚钱?)或在图右侧有大量数据(这种情况往往发生在变量设置上限的时候，比如说将超过 500 000 美元的房屋都计为 500 000 美元)，就该引起警觉了。

箱形图[①]可以用来比较多组数据，图 5.2 中比较了不同质量评分的房屋售价数据，其中 1 代表质量极差，10 代表质量极佳。

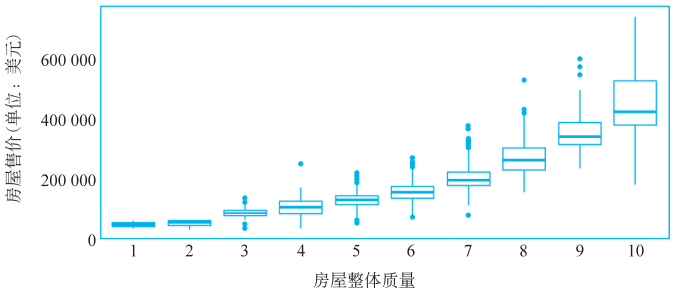

图 5.2 用箱形图比较不同质量房屋的售价

① 箱形图又称盒式图，图中的方形代表了中间的一半数据(也就是 0.25 分位数到 0.75 分位数之间的数据)，方形中间的横线表示中位数，两侧的线段显示了剩余数据的范围。线段以外的点用来表示可能的离群值。

在图 5.2 中，房屋售价与房屋整体质量之间的关系符合人们的直觉，因为高质量房屋往往会以更高的价格卖出。可以看到有一座整体质量为 10 分的房屋卖出了 200 000 美元(见下方线段的终点)，但我们有理由认为，有其他因素导致了它比其他 10 分房屋的售价更低。这是数据工作者应该特别注意的信息。

图 5.3 中的条形图也可以用来显示分类数据的分布。

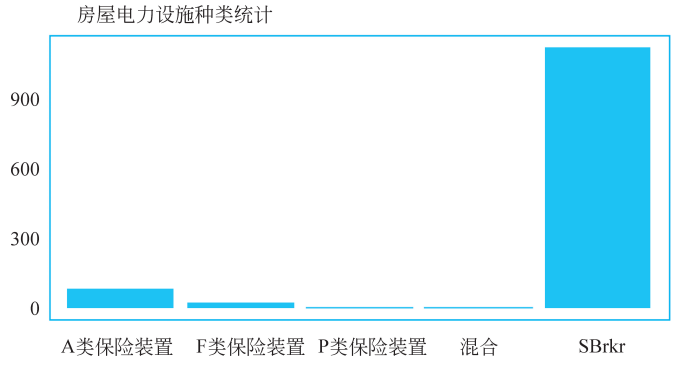

图 5.3　电力设施种类条形图

并非所有的图表都值得一看，但检查图表仍然有助于回答前面的问题：变量取值是否符合直觉？图 5.3 显示，绝大多数的房屋都采用了同一种电力设施，对于当前的任务，这是一项有意义的信息：由于这个变量的取值几乎完全一致，那么它就不会显著影响房屋的售价。

图 5.4 则展示了售出房屋数量随月份的变化，你可以很明显地从中看到每年夏季是房屋售卖高峰，而冬季则是低谷。这种现象称为**季节性**(seasonality)。折线图可以帮助你发现这类趋势。

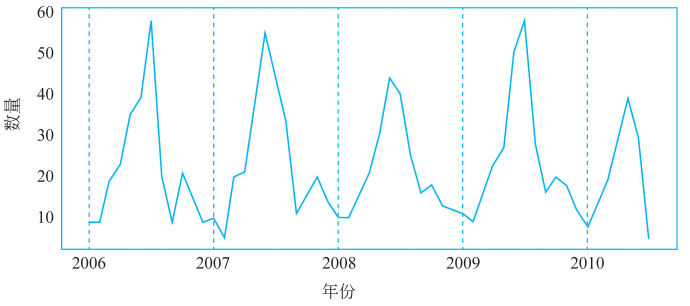

图 5.4 售出房屋数量随时间变化的折线图

图 5.5 是房屋售价与面积(如果是多层住宅,取单层的面积)之间的散点图。图 5.5 中的趋势也符合我们的直觉,一般而言,更大的房子就会更贵。当然,也会有例外,有的时候小房子反而比大房子更贵。其他因素会对房屋售价造成影响,但整体趋势是符合我们直觉的。而当我们想要预测房屋售价时,房屋面积看上去可以提供很有价值的参考。

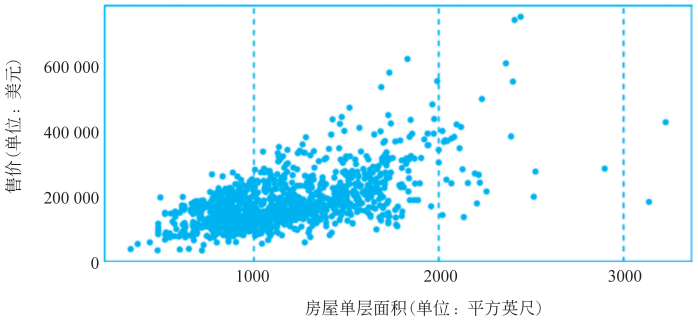

图 5.5 房屋售价与面积的散点图

这一部分简要介绍了你能快速从图表当中获取的信息。如果你想更深入地了解数据可视化,我们在此推荐两本书:

- *Now You See it*：*Simple Visualization Techniques for Quantitative Analysis*，Stephen Few（Analytics Press，2009）
- *The Visual Display of Quantitative Information*，Edward Tufte(Graphics Press，2011)

注意：离群值与缺失数据

每个数据集当中都会有异常值、离群值和缺失数据，重要的是如何处理这些状况。

比如说，在图 5.2 中，有几个数据点被标记为可能的离群值。但不要仅因为将这些点划为了离群值，就不假思索地认为它们毫无用处，可以直接删掉。如果仅因为可视化结果就删除有用的数据点，那么你的预测软件将永远无法超过 Zillow 公司。相反，应该考虑问题的背景：在房地产行业，常常会出现某间房屋的售价远高于其他房屋的情况。回想第 4 章中的内容，在移除离群值之前，你需要给出充分的商业理由。在这个例子中，你能找到合适的理由吗？

除离群值外，缺失数据呢？如果“地下室面积”数据缺失，是代表这间房屋有地下室而面积未知，还是意味着没有地下室，因此应该被设为 0？

如果前面的内容看起来过于琐碎，那其实是笔者有意为之。数据工作者在从事一个项目时，需要做出数以百计的小型决策，而它们积累起来，影响将会是巨大的。如果任由数据工作者自行探索，而不为他们提供该领域内的专业意见，那么数据工作者可能会不断地削减数据，移除困难的数据点，直到数据本身无法