第 3 章　数据分析师岗位情况分析

——数据预处理

使　命

①了解在数据分析之前如何进行数据预处理；②完成“数据分析师岗位情况分析”案例的第二步——数据预处理；③提升批判性地使用 AI 工具的意识；④沉浸于“问题逻辑认知模式”中，通过 5E 路径围绕解决问题开展学习。

3.1 Excitation——提出问题

我们在第 2 章已经获取到 450 条招聘数据分析师岗位的数据。在进行数据分析之前，首先了解这个数据集。图 3-1 是用 Excel 打开数据集文件“lagou_data.csv”后看到的部分内容。

Column1	positionNa	city	district	salary	workYear	education	companyName	industry	financeStage	companySize	companyLabelList	positionAdvantage
0	数据分析师	深圳	西丽	25k-40k	经验1-3年	本科	雄岸區塊鏈集團	区块链	上市公司	150-500人	R语言,科技金融,金融业	“高薪高福利”
1	数据分析师	广州	珠江新城	15k-30k	经验1-3年	本科	广发银行信用卡中心	金融业	不需要融资	2000人以上	PL/SQL,MySQL,金融业,	“大平台、大团队”
2	数据分析师	深圳	南山区	15k-30k	经验3-5年	本科	富途	科技金融	上市公司	500-2000人	R语言,科技金融,Pythor	“能够熟悉券商的内部运作机制，了解证券业务”
3	数据分析师	北京	海淀区	20k-40k	经验3-5年	本科	字节跳动	内容资讯,3	D轮及以上	2000人以上	弹性工作,就近租房补贴	“六险一金,弹性工作,租房补贴,免费三餐”
4	数据分析师	北京	三元桥	15k-20k	经验1-3年	本科	欢喜传媒	影视｜动漫	上市公司	50-150人	视频,产品策划,需求分析	“互联网公司”
5	数据分析师	上海	天山路	25k-50k	经验3-5年	本科	拼多多	电商平台	上市公司	2000人以上	电商平台,SQL	“大平台”
6	数据分析师	深圳	宝安区	15k-30k	经验1-3年	本科	字节跳动	内容资讯,3	D轮及以上	2000人以上	弹性工作,就近租房补贴	“六险一金,弹性工作,免费三餐,餐补”
7	数据分析师	上海	静安区	8k-15k	经验1-3年	本科	亿品	营销服务	不需要融资	150-500人	电商平台,SQL	“双休不加班，团队氛围好，技术提升快”
8	数据分析师	上海	天山路	20k-40k	经验3-5年	本科	拼多多	电商平台	上市公司	2000人以上	电商平台,Python	“一年两次调薪机会 团队氛围佳”
9	数据分析师	北京	白石桥	20k-40k	经验3-5年	本科	易车公司	汽车交易与	上市公司	2000人以上	汽车交易平台,Python	“发展前景好”
10	数据分析师	上海	北外滩	20k-39k	经验3-5年	本科	众安科技	企业服务,3	不需要融资	500-2000人	科技金融,SQL,Python	“前景”

图 3-1　数据集文件的文件头和前面部分数据

该数据集共有 13 个字段，450 条数据，每个字段名称由英文标注。表 3-1 是字段的含义。

表 3-1　数据集各字段含义

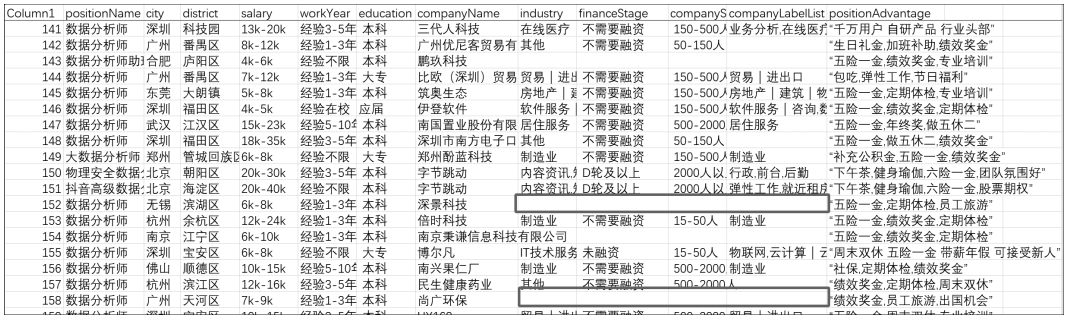
数据集字段名称	含　义
Column1	序号
positionName	职位名称
city	城市
district	区域

续表

数据集字段名称	含　义
salary	薪资
workYear	工作经验
education	学历要求
companyName	企业名称
industry	企业所属行业
financeStage	融资阶段
companySize	企业规模
companyLabelList	企业标签
positionAdvantage	福利待遇

通过浏览这些数据，我们发现爬取的这些数据还存在以下明显的问题。

（1）数据不完整（数据缺失）。如图 3-2 中缺少 industry、financeStage、companySize、companyLabelList 等数据。



Column1	positionName	city	district	salary	workYear	education	companyName	industry	financeStage	companyS	companyLabelList	positionAdvantage
141	数据分析师	深圳	科技园	13k-20k	经验3-5年	本科	三代人科技	在线医疗	不需要融资	150-500人	业务分析,在线医疗	"千万用户 自研产品 行业头部"
142	数据分析师	广州	番禺区	8k-12k	经验1-3年	本科	广州优尼客贸易有	其他	不需要融资	50-150人		"生日礼金,加班补助,绩效奖金"
143	数据分析师助	合肥	庐阳区	4k-6k	经验不限	本科	鹏玖科技					"五险一金,绩效奖金,专业培训"
144	数据分析师	广州	番禺区	7k-12k	经验1-3年	大专	比欧（深圳）贸易	贸易｜进出	不需要融资	150-500人	贸易｜进出口	"包吃,弹性工作,节日福利"
145	数据分析师	东莞	大朗镇	5k-8k	经验1-3年	本科	筑奥生态	房地产｜装	不需要融资	150-500人	房地产｜建筑｜物	"五险一金,定期体检,专业培训"
146	数据分析师	深圳	福田区	4k-5k	经验在校	应届	伊登软件	软件服务｜	不需要融资	150-500人	软件服务｜咨询,客	"五险一金,绩效奖金,定期体检"
147	数据分析师	武汉	江汉区	15k-23k	经验5-10年	本科	南国置业股份有限	居住服务	不需要融资	500-2000	居住服务	"五险一金,年终奖,做五休二"
148	数据分析师	深圳	福田区	18k-35k	经验3-5年	本科	深圳市南方电子口	其他	不需要融资	50-150人		"五险一金,做五休二,绩效奖金"
149	大数据分析师	郑州	管城回族区	6k-8k	经验不限	大专	郑州酚蓝科技	制造业	不需要融资	150-500人	制造业	"补充公积金,五险一金,绩效奖金"
150	物理安全数据	北京	朝阳区	20k-30k	经验3-5年	本科	字节跳动	内容资讯,D	轮及以上	2000人以	行政,前台,后勤	"下午茶,健身瑜伽,六险一金,团队氛围好"
151	抖音高级数据	北京	海淀区	20k-40k	经验不限	本科	字节跳动	内容资讯,D	轮及以上	2000人以	弹性工作,就近租房	"下午茶,健身瑜伽,六险一金,股票期权"
152	数据分析师	无锡	滨湖区	6k-8k	经验1-3年	本科	深景科技					"五险一金,定期体检,员工旅游"
153	数据分析师	杭州	余杭区	12k-24k	经验1-3年	本科	倍时科技	制造业	不需要融资	15-50人	制造业	"五险一金,绩效奖金,定期体检"
154	数据分析师	南京	江宁区	6k-10k	经验1-3年	本科	南京秉谦信息科技	有限公司				"五险一金,绩效奖金,定期体检"
155	数据分析师	深圳	宝安区	6k-8k	经验不限	大专	博尔凡	IT技术服务	未融资	15-50人	物联网,云计算｜云	"周末双休 五险一金 带薪年假 可接受新人"
156	数据分析师	佛山	顺德区	10k-15k	经验5-10年	本科	南兴果仁厂	制造业	不需要融资	500-2000	制造业	"社保,定期体检,绩效奖金"
157	数据分析师	杭州	滨江区	12k-16k	经验3-5年	本科	民生健康药业	其他	不需要融资	500-2000人		"绩效奖金,定期体检,周末双休"
158	数据分析师	广州	天河区	7k-9k	经验1-3年	本科	尚广环保					"绩效奖金,员工旅游,出国机会"

图 3-2　数据集中的数据不完整问题

（2）数据相同（数据重复）。如图 3-3 所示，序号第 44 行和序号第 47 行数据完全相同。

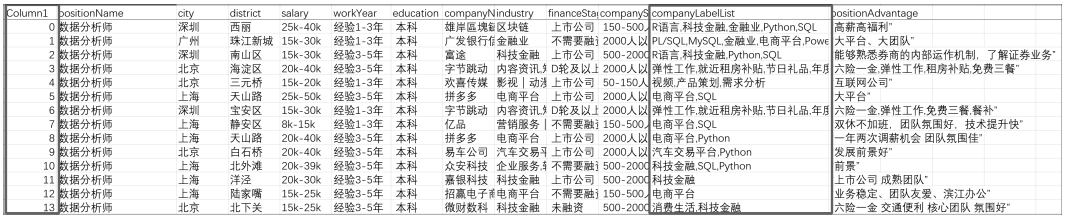
Column1	positionName	city	district	salary	workYear	education	companyN	industry	financeSta	companySi	companyLabelList	positionAdvantage
42	高级数据分析师	上海	洋泾	30k-45k	经验5-10年	本科	奇富科技	科技金融	上市公司	500-2000	Excel,科技金融,SQL	"发展空间大 团队氛围好"
43	资深数据分析师（北京）	北京	望京	30k-40k	经验5-10年	硕士	百融云创	科技金融,人	上市公司	500-2000	PL/SQL,R语言,科技金融,风控建模,决策	"上市公司"
44	商业化数据分析师	上海	凉城	20k-40k	经验1-3年	本科	肯斯爪特	游戏	C轮	50-150人	社交平台,Python	"团队氛围好，晋升空间大"
45	数据分析师（智能客服）	上海	长宁区	25k-50k	经验3-5年	本科	拼多多	电商平台	上市公司	2000人以	BI,SQL,数据库	"平台大,大牛多,氛围好"
46	高级数据分析师（北京）	北京	望京	20k-30k	经验3-5年	硕士	百融云创	科技金融,人	上市公司	500-2000	科技金融,风控建模,数据建模,SQL	"上市公司"
47	商业化数据分析师	上海	凉城	20k-40k	经验1-3年	本科	肯斯爪特	游戏	C轮	50-150人	社交平台,Python	"团队氛围好，晋升空间大"
48	高级数据分析师（北京）	北京	朝阳区	25k-35k	经验3-5年	硕士	百融云创	科技金融,人	上市公司	500-2000	人工智能服务,SAS,科技金融,逻辑回归	"成长空间大，平台稳定"
49	BI业务数据分析师	上海	洋泾	18k-30k	经验3-5年	本科	奇富科技	科技金融	上市公司	500-2000	EXCEL,科技金融,SQL	"公司发展好 职业发展空间大"
50	Bigo live-业务数据分析师	广州	番禺区	10k-15k	经验不限	本科	BIGO	社交媒体	上市公司	2000人以	MCN｜直播平台,社交平台,SQL	"国际化视野；领先直播app；优秀团队伙伴"
51	游戏数据分析师	北京	海淀区	20k-40k	经验3-5年	本科	边锋	游戏	不需要融资	500-2000	回归算法	"团队发展好"
52	投放数据分析师	上海	虹口区	15k-30k	经验不限	本科	肯斯爪特	游戏	C轮	50-150人	社交平台,SQL	"团队氛围好，晋升空间大"

图 3-3　数据集中的数据重复问题

（3）存在与分析的问题无关的数据。如图 3-4，表中的 Column1 和 companyLabelList 与分析的问题无关联。

（4）数据取值异常。如图 3-5，表中的 Column1 取值 231 所在的这行数据，该行数据中的 salary 薪资取值远大于其他薪资。

（5）目前的数据无法直接用于分析。例如，数据集中的“薪资”和“工作经验”字段的取值是一个范围，不是一个具体的数值（见图 3-5），无法进行后续的相关分析，如工作经验与



Column1	positionName	city	district	salary	workYear	education	companyNa	industry	financeSta	companyS	companyLabelList	positionAdvantage
0	数据分析师	深圳	西丽	25k-40k	经验1-3年	本科	雄岸區塊鏈	区块链	上市公司	150-500人	R语言,科技金融,金融业,Python,SQL	高薪高福利"
1	数据分析师	广州	珠江新城	15k-30k	经验1-3年	本科	广发银行信	金融业	不需要融资	2000人以上	PL/SQL,MySQL,金融业,电商平台,Powe	大平台、大团队"
2	数据分析师	深圳	南山区	15k-30k	经验3-5年	本科	富途	科技金融	上市公司	500-2000	R语言,科技金融,Python,SQL	能够熟悉券商的内部运作机制，了解证券业务"
3	数据分析师	北京	海淀区	20k-40k	经验3-5年	本科	字节跳动	内容资讯,D	D轮及以上	2000人以	弹性工作,就近租房补贴,节日礼品,年月	六险一金,弹性工作,租房补贴,免费三餐"
4	数据分析师	北京	三元桥	15k-20k	经验1-3年	本科	欢喜传媒	影视｜动漫	上市公司	50-150人	视频,产品策划,需求分析	互联网公司"
5	数据分析师	上海	天山路	25k-50k	经验3-5年	本科	拼多多	电商平台	上市公司	2000人以	电商平台,SQL	大平台"
6	数据分析师	深圳	宝安区	15k-30k	经验1-3年	本科	字节跳动	内容资讯,D	D轮及以上	2000人以	弹性工作,就近租房补贴,节日礼品,年月	六险一金,弹性工作,免费三餐,餐补"
7	数据分析师	上海	静安区	8k-15k	经验1-3年	本科	亿品	营销服务｜	不需要融资	150-500人	电商平台,SQL	双休不加班，团队氛围好，技术提升快"
8	数据分析师	上海	天山路	20k-40k	经验3-5年	本科	拼多多	电商平台	上市公司	2000人以	电商平台,Python	一年两次调薪机会 团队氛围佳"
9	数据分析师	北京	白石桥	20k-40k	经验3-5年	本科	易车公司	汽车交易平	上市公司	2000人以	汽车交易平台,Python	发展前景好"
10	数据分析师	上海	北外滩	20k-39k	经验3-5年	本科	众安科技	企业服务,金	不需要融资	500-2000	科技金融,SQL,Python	前景"
11	数据分析师	上海	洋泾	20k-30k	经验3-5年	本科	嘉银科技	科技金融	上市公司	500-2000	科技金融	上市公司 成熟团队"
12	数据分析师	上海	陆家嘴	15k-25k	经验3-5年	本科	招赢电子商	电商平台	不需要融资	150-500人	电商平台	业务稳定、团队友爱、滨江办公"
13	数据分析师	北京	北下关	15k-25k	经验3-5年	本科	微财数科	科技金融	未融资	500-2000	消费生活,科技金融	六险一金 交通便利 核心团队 氛围好"

图 3-4　数据集中与后续分析无关的数据列

Column1	positionName	city	district	salary	workYear	education	companyName	industry	financeStage	companySize	companyLabelList	positionAdvantage
228	数据分析师	无锡	锡山区	15k-30k	经验3-5年	本科	无锡黑骑士科技					"定期体检,工作餐,出差补贴"
229	数据分析师	上海	浦东新区	10k-15k	经验1-3年	本科	立邦	能源｜矿产｜环保	不需要融资	2000人以上	能源｜矿产｜环保"五险一金,定期体检,交通补贴"	
230	数据分析师	深圳	福田区	15k-23k	经验3-5年	本科	比亚迪	制造业	不需要融资	2000人以上	制造业	"免费班车,五险一金,绩效奖金"
231	数据分析师	兰州	安宁区	40k-80k	经验3-5年	大专	金瑞鸿盛	数据服务｜咨询	未融资	50-150人	软件服务｜咨询	"时间扁平化"
232	数据分析师	兰州	安宁区	6k-10k	经验3-5年	大专	金瑞鸿盛	数据服务｜咨询	未融资	50-150人	数据服务｜咨询,影	"适合社恐人群"
233	数据分析师	广州	海珠区	6k-10k	经验1-3年	本科	美业颜究院	数据服务｜咨询,营销	B轮	50-150人	Excel,数据服务｜咨	"五险一金，带薪年假"
234	数据分析师	深圳	宝安区	10k-15k	经验在校	应届	京信数科					"五险一金,周末双休,餐饮补贴"
235	数据分析师	宜春	樟树市	5k-8k	经验1-3年	本科	仁和药业股份有限	医疗｜保健｜美容	不需要融资	2000人以上	医疗｜保健｜美容"五险一金,绩效奖金,员工旅游"	

图 3-5　数据集中数据取值异常

职位需求、工作经验与薪资、薪资与学历、不同城市薪资待遇分析等。

可见，我们首先需要对这个数据集进行必要的处理才能进行后面的分析。

那么，在数据分析之前，如果数据集还存在问题，我们应该如何处理呢？

3.2 Exploration——探索问题本质

3.2.1 AI 助学

如果我们遇到了上面的问题不知道如何下手，不妨借助 AI 工具找到解决问题的方向。

问题：

在进行数据分析时，数据还存在数据不全、数据异常、数据重复、有的数据无法直接用于分析等问题，应该如何处理？

文心一言回答：

在进行数据分析时，遇到数据不全、数据异常或某些数据无法进行分析的问题，需要采取一系列的处理方法，以确保分析的有效性和准确性。以下是一些建议的步骤和方法。

（1）数据清洗。

- 处理缺失值。根据缺失值的类型和分布情况，可以选择删除含有缺失值的记录、使用均值、中位数、众数或其他统计方法进行填充，或者使用插值、预测等方法进行估算。
- 处理异常值。异常值可能是数据输入错误、测量误差或其他原因导致的。可以通过统计分析、绘制箱线图、使用 3σ 原则等方法识别异常值，并根据具体情况选择删除、替换或保留这些值。

（2）数据转换。

- 对于不符合分析要求的数据，如格式不正确、单位不统一等，需要进行数据转换。例如，将文本格式的数字转换为数字格式，将不同单位的数据转换为统一单位等。
- 对于非线性关系的数据，可以考虑进行对数转换、Box-Cox 转换等，以改善数据的分布和线性关系。

（3）数据插补。

对于缺失值，除了上述的填充方法外，还可以使用插补技术，如 K 近邻插补、多重插补等，以更精确地估算缺失值。

(4) 数据降维。

对于高维数据，可以考虑使用主成分分析(PCA)、因子分析等方法进行降维，以减少计算量并提高分析的准确性。

(5) 数据合并与整合。

如果数据来源多个渠道或表格，需要进行数据合并与整合。在合并数据时，需要注意数据的一致性和完整性，避免重复和遗漏。

(6) 数据筛选。

根据分析目的和需要，对数据进行筛选，选择相关的变量和记录进行分析。同时，要注意去除无关紧要的变量和记录，以提高分析的效率和准确性。

(7) 使用适当的统计方法或模型。

根据数据的类型和分布特点，选择合适的统计方法或模型进行分析。例如，对于连续变量可以使用 t 检验、方差分析等方法；对于分类变量可以使用卡方检验、逻辑回归等方法。

(8) 结果验证与解释。

对分析结果进行验证和解释，确保结果的可靠性和准确性。如果结果不符合预期或存在疑问，需要重新检查数据处理和分析过程，找出问题所在并进行修正。

总之，在进行数据分析时，遇到数据不全、数据异常或某些数据无法进行分析的问题时，需要采取一系列的处理方法和技术手段，以确保数据的质量和准确性，从而得到可靠的分析结果。

从 AI 工具的回答中可以看到，在进行数据分析之前，可能需要对数据进行一系列处理。同时，AI 工具给出的(7)和(8)与所提的问题无关，应该是数据分析的方法和结果验证的内容，将其忽略即可。

3.2.2 探索问题本质——找准靶心

在进行数据分析之前，一般需要对将要分析的数据进行处理，这个过程称为**数据预处理**。在 3.1 节中，我们浏览数据发现的问题包括数据缺失、数据重复、数据异常、存在与分析无关的字段以及数据无法直接用于分析等。基于 AI 的帮助，我们将要进行的数据预处理任务分为以下三个子任务。

(1) 数据清洗。

发现并处理数据集中存在的缺失值、重复值和异常值。

(2) 特征选择。

从数据集中选择对分析目标有用的字段(特征)，从而提高处理的效率，增强分析结果可解释性。例如，数据中的“Column1”和“companyLabelList”等字段与我们的分析目标无关，可以直接去除。

(3) 数据转换。

将原始不能直接用于分析的数据进行转换，使其满足数据分析的需求。例如，我们数据中的薪资和工作经验字段等需要转换成数值。

3.3 Enhancement——拓展求解问题必备的知识和能力

3.3.1 选择数据预处理工具

明确了数据预处理要完成的具体任务后，接下来是如何进行数据预处理。首先，我们想到的就是使用工具。常见的数据预处理工具有哪些呢？这里，我们可借助 AI 工具了解。

问题：

常见的数据预处理工具有哪些？

文心一言回答：

常用的数据预处理工具包括以下几类。

(1) 编程语言库。

- Pandas。Python 中的一个强大的数据分析工具，提供了数据结构和数据处理功能，如 Series、DataFrame 等，方便进行数据清洗和转换。
- NumPy。Python 中用于处理数组和矩阵运算的库，提供了数学函数和线性代数运算功能，可用于数据的数值变换。
- SciPy。Python 中用于科学计算的库，提供了矩阵支持、数值计算模块等功能，可用于数据处理和分析。
- R 语言。R 是统计分析领域的常用语言，拥有大量的数据处理和分析包，如 tidyverse、dplyr 等，方便进行数据清洗和转换。

(2) 数据清洗工具。

- OpenRefine(原名 Google Refine)。一个开源的数据清洗工具，通过可视化界面和交互式工具帮助识别、清理和转换数据集中的错误、重复或不完整的数据。
- Trifacta Wrangler。一个自动化的数据清洗工具，提供了数据探查、数据清洗和数据准备等功能，帮助用户快速清洗和整理数据。

(3) ETL 工具。

- Talend。一个数据集成和数据质量工具，提供了数据清洗、数据转换和数据匹配等功能，确保数据的一致性和准确性。
- Apache Nifi。一个易于使用、功能强大的数据流工具，可以帮助设计、构建和管理实时数据流，包括数据清洗和转换等任务。

(4) 数据可视化工具。

- Tableau。一款强大的数据可视化工具，提供了丰富的图表类型和交互功能，帮助用户更好地理解和分析数据。
- Power BI。微软推出的商业智能工具，提供了数据预处理、数据可视化和数据分析等功能，方便用户进行数据探索和分析。

这些工具各有特点和优势，可以根据具体的数据情况和需求选择合适的工具进行数据预处理。同时，也可以结合多种工具的功能进行组合使用，以满足复杂的数据处理需求。

在 AI 的帮助下，我们知道了 Python 的 Pandas 是一个强大的数据分析工具，方便进行

数据清洗和转换。因此决定使用 Pandas 作为数据预处理工具，完成数据清洗、特征选择和数据转换等任务。

提示： Pandas 是 Python 中专门用于数据处理和数据分析的第三方库，需要单独下载并安装。

3.3.2 Pandas 基础

使用 Pandas 进行数据预处理之前，需要先学习关于数据与 Pandas 的一些相关概念和基础知识。

3.3.2.1 数据的相关概念

1. 数据元素

数据元素，也称为结点或记录。一个数据元素可由若干数据项（属性、字段）组成。例如，表 3-2 所示的考生数据，每名考生的数据就是一个数据元素，包括准考证号、姓名和语文、数学、英语、历史、地理、政治 6 门课程的成绩。

2. 一维数据

一维数据是指数据元素由一个因素即可确定。例如，一名考生的数据就是一维数据，如表 3-2 所示。表中的阴影部分就是一维数据，它的某个数据项由该数据项所在位置这一个因素便可确定。例如，姓名“张珊”这个数据项仅由它所在的位置 2 即可确定。

表 3-2 一维数据示例

含义	准考证号	姓名	语文	数学	英语	历史	地理	政治
位置	1	2	3	4	5	6	7	8
数据	KS00001	张珊	109	120	114	78	82	90

3. 二维数据

二维数据由多个一维数据组成，二维数据中的某项数据需要由两个因素共同决定。例如，多名考生的数据就是一个二维数据，如表 3-3 所示。表中的阴影部分就是一个二维数据，它的某一个数据项需要由该数据项所在的行和列两个因素才能确定。例如，李思的数学成绩 116 这个数据项，需要由该数据项所在的行号 2 和所在的列号 4 两个因素共同确定。

表 3-3 二维数据示例

含义	准考证号	姓名	语文	数学	英语	历史	地理	政治
行号	列　号							
	1	2	3	4	5	6	7	8
1	KS00001	张珊	109	120	114	78	82	90
2	KS00002	李思	123	116	137	84	92	94
3	KS00003	王武	133	129	132	83	85	78

3.3.2.2　Pandas 的数据结构

Pandas 提供了 **Series** 和 **DataFrame** 两种数据结构，分别用于处理一维数据和二维数据。这两种数据结构能够满足处理金融、统计、社会科学、工程等领域里绝大部分问题的需求。

1. Series

Series 是一组有索引标签且数据类型相同的一维数据结构。Series 对象包括两个部分：①values，一组数据；②index，相关数据的索引标签。

【例 3-1】　基于列表数据创建一个 Series 对象。

```
1   #加载 Pandas 库
2   import pandas as pd #pd 是 pandas 的别名
3   #由列表创建 Series 对象
4   names=["张珊","李思","王武","李明","徐晓丽"] #定义一个名为 names 的列表
5   print("原始列表为:")
6   print(names)
7   Snames=pd.Series(names)                    #由 names 列表创建一个 Series 对象
8   print("由列表创建的 Series 为:")
9   print(Snames)
```

在上述代码中，第 2 行使用 import 将 pandas 加载到程序中，并使用 as 给 pandas 起了一个别名 pd(也可以是其他名称)。这样，程序中用到 pandas 的地方就可以换成 pd，从而简化程序的书写过程。第 7 行使用 pd.Series()方法创建了一个 Series 对象，该对象中保存的数据来自列表 names 中存储的 5 个姓名字符串。

程序的运行结果如图 3-6 所示。names 列表只包含 5 个字符串，而由列表创建的 Series 除了包含原始列表中的数据值外，还包含标签(索引)信息。

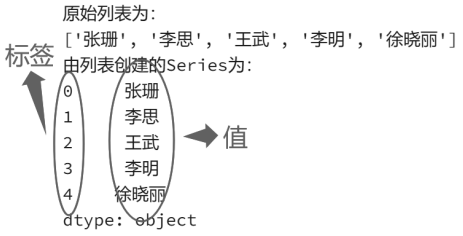


图 3-6　例 3-1 的程序运行结果

2. DataFrame

DataFrame 是具有行标签和列标签的二维表格型数据结构，与 Excel 表类似，每列数据可以是不同的值类型(数字、字符串、布尔型等)。图 3-7 是 DataFrame 结构示意图。

DataFrame 是一个抽象的类，要使用 DataFrame 中的方法，就需要创建一个具体的 DataFrame 对象。创建 DataFrame 对象的语法如下：

```
pandas.DataFrame(data, index, columns, …)
```

	准考证号	姓名	语文	数学	英语	历史	地理	政治
0	KS00001	张珊	109	120	114	78	82	90
1	KS00002	李思	123	116	137	84	92	94
2	KS00003	王武	133	129	132	83	85	78
3	KS00004	李明	142	125	145	89	95	93
4	KS00005	徐晓丽	98	102	119	75	69	81

图 3-7　DataFrame 结构示意图

其中，主要参数的含义：

① data，是所创建的 DataFrame 对象中数据的来源，可以是列表、字典、Series 或 DataFrame 对象等。

② index，是 DataFrame 对象数据的行标签，如果未指定，则默认为 RangeIndex，即(0，1，2，…，$n-1$)，n 为行数。

③ columns，是列标签，如果未指定则默认为 RangeIndex，即(0，1，2，…，$m-1$)，m 为列数。

【例 3-2】　基于列表数据创建一个 DataFrame 对象。

```
1   import pandas as pd                    #pd 是 pandas 的别名
2
3   employees=[
4       ['1001','张伟','男','财务部','1999-7-1'],
5       ['1002','李晓欣','女','审计部','2007-4-5'],
6       ['1003','邓朝阳','男','销售部','2004-2-28']
7   ]
8   #创建 DataFrame 对象 df
9   df = pd.DataFrame(employees,columns=['员工编号','姓名','性别', '部门', '入职
    日期'])
10  #输出 DataFrame 对象 df
11  df
```

程序的运行结果如图 3-8 所示。

	员工编号	姓名	性别	部门	入职日期
0	1001	张伟	男	财务部	1999-7-1
1	1002	李晓欣	女	审计部	2007-4-5
2	1003	邓朝阳	男	销售部	2004-2-28

图 3-8　例 3-2 的程序运行结果

在图 3-8 中，最左侧的 0、1、2 是默认的行标题(标签)；最上面的员工编号、姓名、性别、部门、入职日期则是创建对象时设置的列标题(标签)。

提示：由于实际应用场景中处理的大多都是二维数据，因此下面重点介绍 DataFrame

结构的相关方法。

3.3.2.3　Pandas 读写文件

Pandas 提供了读取 Excel 和 CSV 文件数据的方法，返回值为 DataFrame 对象；还提供了将 DataFrame 数据保存到 Excel 和 CSV 文件中的方法。下面仅介绍 Pandas 读写 CSV 文件的相关方法。读写 Excel 文件的方法与 CSV 类似，读者可自行查阅相关资料。

1. 写文件

DataFrame 的 to_csv()方法可以将 DataFrame 对象的数据写入 CSV 文件。该方法可以设置很多参数，除了必须给出写入文件的路径和名称外，其他参数都可以缺省。例如，可以通过 columns 参数指定写入的列标题；通过 index 参数指定是否写入行标签，默认为 True，表示写入行标签；通过 encoding 参数指定文件使用的字符集编码类型，常见的编码格式包括 UTF-8、GBK、GB2312 等。此处不再一一展开，想深入了解的读者可自行查阅相关资料。

【例 3-3】 使用 Pandas 将例 3-2 中 DataFrame 对象 df 中保存的 3 条员工数据写入名为“员工数据.csv”文件中。

使用 Pandas 的 to_csv()方法非常简单，只需在例 3-2 的代码最后添加如下的一行代码即可。

```
df.to_csv('d:/project/员工数据.csv')
```

其中，“d:/project/”表示写入的文件路径，“员工数据.csv”表示要写入数据的文件名。程序运行完毕，会在所设置的路径下生成该文件。可以使用记事本、Excel 等软件打开查看该文件。

提示： 写入文件时，写入的文件路径须真实存在，否则会报找不到文件“FileNotFoundError”的错误。

2. 读文件

Pandas 的 read_csv()方法用于读取 csv 格式的文件，并将存储文件数据的 DataFrame 对象返回。该方法也可以设置很多参数，除了必须设置读取的文件路径和名称外，其他参数都可以缺省。例如，可以通过 header 参数指定第几行作为列名，默认为 0，表示第一行；通过 sep 参数设置所读取文件的分隔符，默认用逗号分隔；encoding 参数指定文件中字符集使用的编码类型等。

【例 3-4】 使用 read_csv()方法读取图 3-9 所示的“华为 Mate60 销售情况.csv”示例文件。该文件共 4 个字段，有 12 条数据，每个字段分别为产品名称、店铺名称、地理位置和产品价格。假设该文件存储在“d:/project”目录下。

程序代码如下：

```
1  import pandas as pd
2  df = pd.read_csv('d:/project/华为 Mate60 销售情况.csv')
3  #输出 DataFrame
4  df
```

产品名称	店铺名称	地理位置	产品价格
华为Mate60	皇家机行	江苏	7048
华为Mate50	海康智能数码专营店	广东	1358
华为Mate60	易购通信		5800
华为Mate60	蜀都通讯	四川	6189
华为Mate60	津门千机惠	天津	6189
华为Mate60	华奥通讯	北京	10588
华为Mate60	麦克先生腾讯员工创业店	北京	6550
华为Mate60	阿布的数码屋	广东	6190
华为Mate60	嘉恒数码商城	北京	6060.2
华为Mate60	新疆恒天通讯官方店	新疆	6490
华为Mate60	华 为 手 机	广东	
华为Mate60	蜀都通讯	四川	6189

图 3-9　华为 Mate60 销售情况.csv 文件的内容示意图

在上述代码中：

第 2 行使用 read_csv()方法读取 CSV 文件的数据，并将数据存储在 DataFrame 对象 df 中，read_csv()方法只给出了第一个参数，代表读取的文件路径和文件名称。

第 4 行是将返回的 df 对象输出到屏幕上。程序的运行结果如图 3-10 所示。

	产品名称	店铺名称	地理位置	产品价格
0	华为Mate60	皇家机行	江苏	7048.0
1	华为Mate50	海康智能数码专营店	广东	1358.0
2	华为Mate60	易购通信	NaN	5800.0
3	华为Mate60	蜀都通讯	四川	6189.0
4	华为Mate60	津门千机惠	天津	6189.0
5	华为Mate60	华奥通讯	北京	10588.0
6	华为Mate60	麦克先生腾讯员工创业店	北京	6550.0
7	华为Mate60	阿布的数码屋	广东	6190.0
8	华为Mate60	嘉恒数码商城	北京	6060.2
9	华为Mate60	新疆恒天通讯官方店	新疆	6490.0
10	华为Mate60	华 为 手 机	广东	NaN
11	华为Mate60	蜀都通讯	四川	6189.0

图 3-10　例 3-4 的程序运行结果

提示：读取文件时，如果遇到编码错误的问题，可以设置 read_csv()方法的 encoding 参数，尝试设置为常见编码 UTF-8、GBK 或 GB2312 等。

3.3.2.4　Pandas 查看数据

DataFrame 查看数据的方法主要包括：

(1) head(n)方法，用于获取前 n 条数据，n 为整数值，缺省时默认为 5。

(2) tail(n)方法，用于获取后 n 条数据，n 为整数值，缺省时默认为 5。

(3) info()方法，用于查看数据的概要信息，包括列数、列标签、列数据类型、每列非空值

数量、内存使用情况等。

(4) describe()方法，用于查看数值型数据的数量、均值、最大值、最小值，方差等信息。

【例 3-5】　使用 info()方法和 describe()方法查看 DataFrame 对象数据的情况。

```
1   #导入 Pandas 库
2   import pandas as pd
3   df = pd.read_csv('d:/project/华为 Mate60 销售情况.csv')
4   df.info()
5   df.describe()
```

在上述代码中：

第 4 行输出的是 DataFrame 对象 df 的概要信息，包括数据条数、数据列、各列的列名、每列非空值的数量、每列的数据类型和占用的内存空间等。

第 5 行输出的是 DataFrame 对象 df 中数值型数据(产品价格)的统计值，包括数量、均值、标准差、最小值、下四分位数、中位数、上四分位数和最大值。程序的运行结果如图 3-11 所示。

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12 entries, 0 to 11
Data columns (total 4 columns):
产品名称    12 non-null object
店铺名称    12 non-null object
地理位置    11 non-null object
产品价格    11 non-null float64
dtypes: float64(1), object(3)
memory usage: 464.0+ bytes
```

	产品价格
count	11.000000
mean	6241.018182
std	2092.527458
min	1358.000000
25%	6124.600000
50%	6189.000000
75%	6520.000000
max	10588.000000

图 3-11　例 3-5 程序的运行结果

3.3.3　用 Pandas 进行数据清洗

3.3.3.1　缺失值检测与处理

缺失值主要包括空值、表示数值缺失的特殊数值(如 NaN、Na、NaT 等)和空字符串等。

Pandas 提供了一系列的方法用来处理缺失数据，主要包括检测缺失值、删除缺失值和填充缺失值。

1. 数据缺失的检测

(1) 空值和 NaN 类缺失值的检测。

Pandas 检测空值或 NaN 类的缺失值时，既可以使用 isnull()方法，也可以使用 isna()方法。这两种方法均可以检测 DataFrame 或 Series 中的缺失值。两种方法返回的是相应位置上的布尔型值，True 表示缺失，False 表示不缺失。

另外，检测时还可以搭配相应统计方法查看数据缺失情况，如搭配 sum()方法可以统计缺失值的数量。

【例 3-6】 检测“华为 Mate60 销售情况.csv”文件中数据是否有空值或缺失值。

```
import pandas as pd
df = pd.read_csv('d:/project/华为 Mate60 销售情况.csv')
#检测是否有缺失值
missing_values = df.isnull()
print("缺失值检测:\n", missing_values)
#统计每列缺失值的数量
missing_count = df.isnull().sum()
print("\n 每列缺失值数量:\n", missing_count)
#统计每行缺失值数量
missing_count_per_row = df.isnull().sum(axis=1)
print("\n 每行缺失值数量:\n", missing_count_per_row)
```

程序的运行结果如图 3-12 所示。

```
缺失值检测:
    产品名称   店铺名称   地理位置   产品价格
0   False  False  False  False
1   False  False  False  False
2   False  False   True  False
3   False  False  False  False
4   False  False  False  False
5   False  False  False  False
6   False  False  False  False
7   False  False  False  False
8   False  False  False  False
9   False  False  False  False
10  False  False  False   True
11  False  False  False  False

每列缺失值数量:
 产品名称    0
店铺名称     0
地理位置     1
产品价格     1
dtype: int64

每行缺失值数量:
 0     0
1      0
2      1
3      0
4      0
5      0
6      0
7      0
8      0
9      0
10     1
11     0
dtype: int64
```

图 3-12 例 3-6 程序的运行情况

由图 3-12 可知，地理位置列数据中存在一个缺失值，位置在第 3 行；产品价格列数据中也存在一个缺失值，位置在第 11 行。

(2) 空字符串检测。

isnull()方法和 isna()方法不适用于检测空字符串，字符串的检测需要用到比较运算符。

【例 3-7】　检测 DataFrame 对象是否有空字符串。

```
1   import pandas as pd
2   #创建一个 DataFrame 对象 df
3   df=pd.DataFrame({'A':[1, 2, ''],'B':[4,'',6]})
4   print('--------------------')
5   print(df)
6   #使用 isnull()方法检测空字符串
7   print('--------------------')
8   print(df.isnull())
9   print('--------------------')
10  #使用比较运算符检测空字符串
11  print(df=="")
```

程序的运行结果如图 3-13 所示。

由图 3-13 可知，isnull()方法并不适用于检测空字符串。

```
--------------------
   A  B
0  1  4
1  2
2     6
--------------------
       A      B
0  False  False
1  False  False
2  False  False
--------------------
       A      B
0  False  False
1  False   True
2   True  False
```

图 3-13　例 3-7 程序的运行结果

2. 删除缺失值

如果某个特征(字段)缺失值比例较高，且这些缺失值对分析目标没有实质性影响，则可以考虑删除缺失值。Pandas 提供的 dropna()方法可以删除用 isnull()方法或 isna()方法检测出的缺失值，其语法格式如下：

```
Pandas.DataFrame.dropna(axis=0, how, * )
```

其中，主要参数的含义：

① axis，用来设置删除的是行还是列，取值为 0 删除行，取值为 1 表示删除列，默认值为 0。

② how，设置删除数据的规则。取值为“all”，表示所有数据均为空时才删除，取值为“any”，表示只要有一个为空就删除，默认取值为“any”。

【例 3-8】　删除“华为 Mate60 销售情况.csv”数据中所有包含缺失值的行。

```
1   import pandas as pd
2   df = pd.read_csv('d:/project/华为 Mate60 销售情况 .csv')
3   #删除所有包含缺失值的行
4   df.dropna()
```

程序的运行结果如图 3-14 所示。

由运行结果可知，dropna()方法删除掉了包含缺失值的第 3 行和第 11 行数据。

提示：这种方法并不能删除包含空字符串的数据。

	产品名称	店铺名称	地理位置	产品价格
0	华为Mate60	皇家机行	江苏	7048.0
1	华为Mate50	海康智能数码专营店	广东	1358.0
3	华为Mate60	蜀都通讯	四川	6189.0
4	华为Mate60	津门千机惠	天津	6189.0
5	华为Mate60	华奥通讯	北京	10588.0
6	华为Mate60	麦克先生腾讯员工创业店	北京	6550.0
7	华为Mate60	阿布的数码屋	广东	6190.0
8	华为Mate60	嘉恒数码商城	北京	6060.2
9	华为Mate60	新疆恒天通讯官方店	新疆	6490.0
11	华为Mate60	蜀都通讯	四川	6189.0

图 3-14　例 3-8 程序的运行结果

3. 填充缺失值

当缺失值的比例相对较低时，可以选择填充缺失值，这样有助于保留数据的完整性，避免因为删除少量数据而损失信息。Pandas 提供的 fillna()方法可以填充用 isnull()或 isna()方法检测出来的缺失值，其语法格式如下：

```
Pandas.DataFrame.fillna(value=None, method=None, axis=None, * )
```

其中，主要参数的含义：

① value，用于填充缺失值的标量值或字典等。

② method，用于设置填充方式。取值为“bfill”表示向后填充，即用缺失值之后的第一个有效值进行填充；取值为“ffill”表示向前填充，即用缺失值之前的第一个有效值进行填充。

该方法返回填充缺失值后的 DataFrame 对象。

对于数值缺失值，可以使用固定值、均值、中位数、插值等进行填充；对于分类特征（即该字段所在列取值为字符串而非数值），可以使用众数、特殊值等进行填充。

众数（Mode）是指在统计分布上具有明显集中趋势点的数值，代表数据的一般水平。也是一组数据中出现次数最多的数值，有时一组数中会有多个众数。

【例 3-9】　填充“华为 Mate60 销售情况.csv”中的缺失值。

```
import pandas as pd
df = pd.read_csv('d:/project/华为 Mate60 销售情况.csv')
#1.使用众数填充分类字段(特征)
#获取'地理位置'列的众数
mode_value = df['地理位置'].mode()[0]
#使用众数填充'地理位置'列中的缺失值,inplace 为 True 表示原地修改
df['地理位置'].fillna(mode_value, inplace=True)
```

```
#2.使用均值填充产品价格
#获取产品价格列的均值
mean = df['产品价格'].mean()
df['产品价格'].fillna(mean,inplace=True)
print(df)
```

在上述代码中：

第 5 行 mode()方法用于获取某列的所有众数，“[0]”用于获取所有众数中频率最高的那个。

第 7 行使用 fillna(mode_value,inplace＝True)方法使用地理位置列的众数填充缺失值，inplace＝True 代表修改原来的 df。

第 10 行 mean()方法用于获取某列的均值。

第 11 行使用产品价格列的均值填充缺失值，并且修改原来的 df。

程序的运行结果如图 3-15 所示。第 3 条数据易购通信缺失的“地理位置”，使用了地理位置列出现次数最多的“北京”进行填充；第 11 条数据缺失的“产品价格”，使用了产品价格列的均值进行填充。

	产品名称	店铺名称	地理位置	产品价格
0	华为Mate60	皇家机行	江苏	7048.000000
1	华为Mate50	海康智能数码专营店	广东	1358.000000
2	华为Mate60	易购通信	北京	5800.000000
3	华为Mate60	蜀都通讯	四川	6189.000000
4	华为Mate60	津门千机惠	天津	6189.000000
5	华为Mate60	华奥通讯	北京	10588.000000
6	华为Mate60	麦克先生腾讯员工创业店	北京	6550.000000
7	华为Mate60	阿布的数码屋	广东	6190.000000
8	华为Mate60	嘉恒数码商城	北京	6060.200000
9	华为Mate60	新疆恒天通讯官方店	新疆	6490.000000
10	华为Mate60	华 为 手 机	广东	6241.018182
11	华为Mate60	蜀都通讯	四川	6189.000000

图 3-15　填充缺失值后的数据

3.3.3.2　重复值检测与处理

Pandas 对重复值的处理主要包括检测重复值和删除重复值。

1. 检测重复值

Pandas 提供的 duplicated()方法可以检测 DataFrame 中的重复数据，其具体语法如下：

```
duplicated(subset=None, keep='first', * )
```

其中，主要参数的含义：

① subset，指定要检查的列标签列表，缺省则检查所有列。

② keep，指定如何标记重复值，默认为“first”，不同取值含义：“first”表示将第一个值视为唯一值，将其余相同的值视为重复值；“last”表示将最后一个值视为唯一值，将其余相

同的值视为重复值；False 表示将所有相同的值均视为重复项。

该方法返回值为布尔型的 Series 结构，True 表示重复，False 表示不重复。

另外，检测重复值时，还可以搭配 sum()方法统计重复数据的数量。

【例 3-10】 使用 duplicated()方法查找重复数据。

```
1  import pandas as pd
2  df = pd.read_csv('d:/project/华为 Mate60 销售情况.csv')
3
4  #检测重复值
5  duplicate_values = df.duplicated()
6  print("重复值检测:\n", duplicate_values)
7
8  #统计重复值数量
9  duplicate_count = df.duplicated().sum()
10 print("\n 重复值数量:", duplicate_count)
```

在上述的代码中：

第 5 行代码使用 duplicated()方法将产品名称、店铺名称、地理位置和产品价格完全相同的记录标记为 True。

第 9 行代码统计重复数据的数量。

程序运行的结果如图 3-16 所示。

```
重复值检测:
 0     False
1      False
2      False
3      False
4      False
5      False
6      False
7      False
8      False
9      False
10     False
11      True
dtype: bool

重复值数量:  1
```

图 3-16 例 3-10 程序的运行结果

2. 删除重复值

Pandas 提供的 drop_duplicates()方法可以删除重复数据，其具体语法如下：

```
DataFrame.drop_duplicates(subset=None,keep='first',
inplace=False,…)
```

其中，sunset 和 keep 参数的含义与 duplicated()方法一致。inplace 参数代表是否原地修改，默认是 False，表示不改变原 DataFrame；True 则表示在原 DataFrame 中删除重复数据。该方法返回的是删除重复值后的 DataFrame。

【例 3-11】 使用 drop_duplicates()删除重复数据。

```
1  import pandas as pd
2  df = pd.read_csv('d:/project/华为 Mate60 销售情况.csv')
3  df_new = df.drop_duplicates()
4  df_new
```

程序运行结果如图 3-17 所示。

以上是处理缺失值和重复值的基本方法。实际场景中，我们根据具体业务的需求，选择适合的处理方式。例如，删除重复值时，可以根据具体需求，选择保留第一个出现的值或最后一个出现的值等。

	产品名称	店铺名称	地理位置	产品价格
0	华为Mate60	皇家机行	江苏	7048.000000
1	华为Mate50	海康智能数码专营店	广东	1358.000000
2	华为Mate60	易购通信	北京	5800.000000
3	华为Mate60	蜀都通讯	四川	6189.000000
4	华为Mate60	津门千机惠	天津	6189.000000
5	华为Mate60	华奥通讯	北京	10588.000000
6	华为Mate60	麦克先生腾讯员工创业店	北京	6550.000000
7	华为Mate60	阿布的数码屋	广东	6190.000000
8	华为Mate60	嘉恒数码商城	北京	6060.200000
9	华为Mate60	新疆恒天通讯官方店	新疆	6490.000000
10	华为Mate60	华 为 手 机	广东	6241.018182

图 3-17　例 3-11 程序的运行结果

3.3.3.3　异常值的检测与处理

异常值(outlier)又称离群点，是指数据中与大多数其他数据明显不同的值。

Pandas 提供了一系列的方法用来处理异常数据，主要包括检测异常值、删除异常值和替换异常值。

1. 检测异常值

常见的判定数据中是否存在异常值的方法包括 3σ 原则和箱线图(箱形图)。

(1) 3σ 原则。

计算数据的均值和标准差，然后根据 3σ 原则，将位于均值±3 倍标准差之外的数据视为异常值。

使用 3σ 原则判断数据集中是否存在异常数据的前提是数据满足正态分布，若数据非正态分布，则可使用箱线图判断异常数据。

如 3.3.2.4 节所述，Pandas 提供的 describe()方法可以计算数据的均值、标准差、上四分位数、中位数、下四分位数等，因此，借助 describe()方法，进行简单计算便可以判定数据集中是否存在异常数据。

(2) 箱线图。

箱线图是一种用作显示一组数据分散情况的统计图，由上限、上四分位数、中位数、下四分位数、下限组成。它最大的优点就是不受异常值的影响，以一种相对稳定的方式描述数据的离散分布情况。通过箱线图，可以直观地发现数据中是否存在异常。

图 3-18 是一个箱线图。根据数据的上四分位数 Q3(分布在 75%的位置)和下四分位数 Q1(分布在 25%的位置)和四分位距(IQR＝Q3－Q1)，将位于下限(Q1－1.5IQR)和上限(Q3＋1.5IQR)之外的数据视为异常值。

为了更直观地查看数据集中的异常数据，可以借助 Pandas 提供的 boxplot()方法绘制

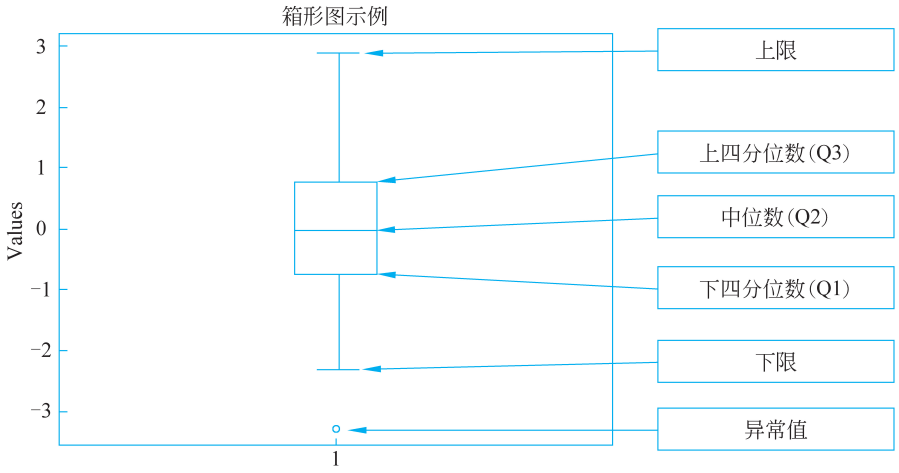


图 3-18 箱线图

箱线图，具体语法如下：

```
DataFrame.boxplot(column=None, * )
```

其中，参数 column 用于指定绘制箱线图的数据列。如果不指定，则默认绘制 DataFrame 中所有数值列的箱线图。

DataFrame 支持按照条件对 DataFrame 中的数据进行筛选。其具体语法如下：

```
df[筛选条件]
```

其中，筛选条件的常见形式为：df['标签']满足某个条件，如 df['标签']等于某个值，大于某个值或小于某个值等，筛选条件返回一个布尔型 Series，通过 df[布尔型 Series]把 DataFrame 中对应标签，布尔值为 True 的数据筛选出来，得到的是 DataFrame 对象。

筛选条件既可以是单个条件，也可以是多个条件，多个条件之间使用 Pandas 的逻辑运算符与(&)、或(|)、非(~)连接。需要注意的是，每个筛选条件需要用小括号括起来。

【例 3-12】 使用箱线图法判断"新-华为 Mate50 销售情况.csv"文件(如图 3-19 所示)中是否存在异常数据。与华为 Mate60 文件结构相同，该文件也有 4 个字段，共 9 条数据，每个字段分别为产品名称、店铺名称、地理位置和产品价格。同样，假设该文件存储在"d:/project"目录下。

产品名称	店铺名称	地理位置	产品价格
华为Mate50	昊玉数码专营店	广东	1488
华为Mate50	中航数码通讯	广东	3148
华为Mate50	华胜通数码	广东	4549
华为Mate50	新通信手机数码	北京	9999
华为Mate50	昊玉数码专营店	广东	1488
华为Mate50	华诚数码商城	山西	3350
华为Mate50	小野优品	江苏	899
华为Mate50	元宇宙电讯	浙江	930
华为Mate50	方圆泽数码优品	陕西	11930

图 3-19 华为 Mate50 销售情况

```
1   import pandas as pd
2   df = pd.read_csv('d:/project/新-华为 Mate50 销售情况.csv')
3   #绘制箱线图
4   df.boxplot(column='产品价格')
5   #使用 describe 方法查看产品价格数据列统计信息
6   stat_data = df['产品价格'].describe()
7   #从统计信息 stat_data 中获取上四分位数和下四分位数
8   Q3 = stat_data.loc['75%']
9   Q1 = stat_data.loc['25%']
10  #计算四分位距
11  IQR = Q3-Q1
12  #计算下限、上限
13  lower_limit = Q1 -1.5 * IQR
14  upper_limit = Q3 + 1.5 * IQR
15  #筛选数据
16  cond = (df['产品价格']<lower_limit) | (df['产品价格']>upper_limit)
17  #筛选数据
18  df_new = df[cond]
19  df_new
```

在上面的代码中：

第 16 行代码定义了在 DataFrame 中筛选数据的条件。

第 18 行代码筛选出了满足条件的数据。

提示：程序运行时，首先显示第 19 行的 df_new，然后才显示第 4 行绘制的箱线图。在不同的平台上运行以上代码需考虑字体库，用以显示中文。

程序运行结果如图 3-20 所示。

	产品名称	店铺名称	地理位置	产品价格
3	华为Mate50	新通信手机数码	北京	9999.0
8	华为Mate50	方圆泽数码优品	陕西	11930.0

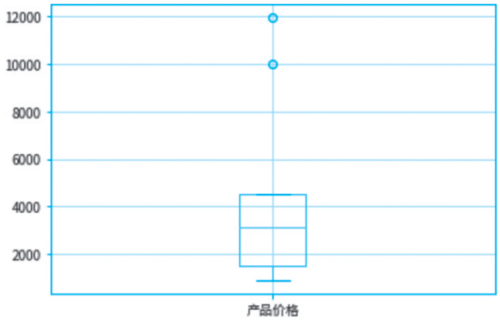


图 3-20　例 3-12 程序的运行结果

由图 3-20 可知，华为 Mate50 产品价格均匀分布在 1500～4500 元。数据中存在两个异常产品价格，即 9999.0 和 11930.0，需要进行适当处理。

2. 删除异常值

如果异常数据在数据集中所占比例很小，可以直接删除异常数据。

删除异常数据有两种方式：一种是对异常数据筛选条件取反，筛选得到的就是不包含异常数据的数据集；另一种方式是通过 drop()方法删除指定索引的数据。下面介绍第一种删除异常的方式。

【例 3-13】 删除新-华为 Mate50 销售情况表中的异常数据。

在例 3-12 下方添加如下两行代码，即可获取到不包含异常值的数据。

```
df_new_1 = df[~cond]
df_new_1
```

程序运行结果如图 3-21 所示。

	产品名称	店铺名称	地理位置	产品价格
0	华为Mate50	昊玉数码专营店	广东	1488.0
1	华为Mate50	中航数码通讯	广东	3148.0
2	华为Mate50	华胜通数码	广东	4549.0
4	华为Mate50	昊玉数码专营店	广东	1488.0
5	华为Mate50	华诚数码商城	山西	3350.0
6	华为Mate50	小野优品	江苏	899.0
7	华为Mate50	元宇宙电讯	浙江	930.0

图 3-21 例 3-13 程序的运行结果

3. 替换异常值

还可以根据需要将异常值替换成特定值，如替换为中位数、均值等，当然也可以通过插值法根据相邻数据点的值推测并替换异常值。

Pandas 提供的 replace()方法可以实现数据替换，具体语法如下：

```
DataFrame.replace(to_replace=None, value=_NoDefault.no_default, *, inplace=
False, *)
```

其中，主要参数的含义：

① to_replace，表示需要被替换的旧值，它可以是一个单一的值，也可以是一个列表或字典，表示多个值要被替换。

② value，表示用于替换的新值，它可以是一个单一的值，也可以是一个列表或字典，表示某些旧值需要被替换为不同的新值。

③ inplace，表示是否直接替换原来的数据，默认值为 False。

【例 3-14】 使用 replace()方法替换新-华为 Mate50 销售数据表中的异常数据。

```
import pandas as pd
df = pd.read_csv('新-华为 Mate50 销售情况.csv')
#产品价格的均值
mean = df['产品价格'].mean()
#产品价格中位数
median = df['产品价格'].median()
#一对一替换
```