



在大模型时代,数据分析的门槛被显著降低。用户通过与大模型进行自然语言交互,无须编写复杂的 SQL 查询或 Python 代码,就能快速获取分析结果、生成可视化图表。本章主要介绍数据分析概述、数据预处理、统计分析、自回归时间序列预测、数据可视化。

5.1 数据分析概述

在“互联网+”时代下,数据正在以超凡的速度呈现爆炸式增长。从移动互联网,到企业私有云及公有云数据平台,再到“感知万物”的物联网等,数据早已经渗透到中国各行各业,数据不再是社会生产的“副产物”,而是可被二次乃至多次加工的原料,从中可以探索更大价值,它变成了生产资料。

数据分析是指用适当的统计分析方法(如聚类分析、相关性分析等),对收集来的大量数据进行分析,从中提取有用信息、形成结论,进而对数据加以详细研究和概括总结的过程。例如,营销人员通过数据分析消费者的购物行为,研究其购物习惯,以便精准地预测下一步的消费;网络安全人员通过实时监测异常的网络活动数据和相应的访问行为,识别出恶意用户进行入侵的模式,做到未雨绸缪;金融从业者可以通过数据分析规避投资风险。

数据分析的目的是从大量的数据中抽取不易推断的信息,所谓信息就是对数据集进行处理,从中提炼出可用于其他场合的结论,可以将从原始数据集中抽取信息的过程称为数据分析的过程。人们一旦理解了这些信息,就能够对产生数据的系统的运行机制进行研究,从而对系统可能的响应和演变作出预测。

数据分析的过程也是数据建模的过程,一旦建立数学模型就可以预测在给定输入的情况下,系统会给出怎样的输出。数据分析的目标不止于建模,更重要的是其预测能力。

5.2 数据预处理

采集到的原始数据通常来自多个异构数据源,数据在准确性、完整性和一致性等方面存在着多种多样的问题。在数据分析之前,首先要做的就是对数据进行预处理,处理数据中的“脏数据”,从而提高数据分析的准确性和有效性。

5.2.1 缺失值处理

数据的缺失主要包括记录的缺失和记录中某个字段信息的缺失,两者都会造成分析结果的不准确。造成数据缺失的原因是多方面的,主要有以下几种。

(1) 有些信息暂时无法获取。例如,在医疗数据库中,并非所有病人的所有临床检验结果都能在给定的时间内得到,致使一部分属性值空缺出来。

(2) 有些信息是被遗漏的。可能是因为输入时认为不重要、忘记填写或对数据理解错误而遗漏,也可能是由于数据采集设备的故障、存储介质的故障、传输媒体的故障、一些人为因素等而丢失。

(3) 有些对象的某个或某些属性本身不适用,如未婚者的配偶姓名、儿童的固定收入状况等。

(4) 有些信息(被认为)是不重要的,如数据库的设计者并不在乎某个属性的取值。

(5) 获取这些信息的代价太大。

对缺失值的处理,要具体问题具体分析,因为属性缺失有时并不意味着数据缺失,缺失本身是包含信息的,所以需要根据不同应用场景下缺失值可能包含的信息,对缺失值进行合理处置。

【例 5-1】 基于提供的产品销售数据,依次执行以下操作,最终以表格形式返回大模型的处理结果。

提示词如下。

请对提供的销售数据表格进行缺失值处理。

(1) 首先计算并显示每列的缺失值百分比。

(2) 对于数值列,缺失值使用同月份的中位数填充。

(3) 对于分类列(如产品类别、地区),缺失值标记为“未知”。

(4) 删除缺失值超过 30% 的列。

(5) 最后输出处理后的完整表格和处理日志。

在 DeepSeek 对话界面单击上传附件图标上传要处理的销售数据表格文件,表格内容如表 5-1 所示。

表 5-1 销售数据表格

A	B	C	D	E	F	G	H	I	J
日期	销售员	产品类别	产品 ID	销售数量	单价(元)	总金额	客户类型	地区	支付方式
2023/1/5	张三	电子产品	P1001	3	1200	3600	企业客户	华东	信用卡
2023/1/7	李四	办公用品	P2005		45		中小企业	华北	现金
2023/1/8	王五	家具	P3002	1	2500	2500		华南	支付宝
2023/1/10	张三	电子产品	P1003	2	800	1600	个人客户		微信
2023/1/12	李四		P2001	5	30	150	中小企业	华东	现金
2023/1/15	王五	家具	P3003	4	1800		企业客户	华北	
2023/1/18	赵六	办公用品	P2007	4	60	240	个人客户	华南	信用卡

续表

A	B	C	D	E	F	G	H	I	J
2023/1/20	张三		P1002	1	1500		企业客户	华东	支付宝
2023/1/22	李四	办公用品	P2003	2	50	100	中小企业		现金
2023/1/25	王五	家具		3	2200	6600	个人客户	华北	微信
2023/1/28	赵六	电子产品	P1004	1	950	950	个人客户	华南	
2023/1/30	张三	办公用品	P2006		75		企业客户	华东	信用卡
2023/2/2	李四	家具	P3004	2		5000	中小企业	华北	支付宝
2023/2/5	王五	电子产品	P1005	4	1100	4400		华南	微信
2023/2/8	赵六		P2008	3	40	120	个人客户	华东	现金
2023/2/10	张三	办公用品	P2009	1	65	65	企业客户		信用卡
2023/2/12	李四	家具	P3005		1950		中小企业	华北	
2023/2/15	王五	电子产品	P1006	2	1350	2700	个人客户	华南	支付宝
2023/2/18	赵六	办公用品	P2010	5		250	企业客户	华东	微信
2023/2/20	钱七	家具	P3006	1	2800	2800	中小企业	华北	现金

大模型回复结果包括以下 5 个方面。

大模型回复

(1) 缺失值百分比统计如下表所示。

表 缺失值百分比统计

列 名	缺失值数量	缺失百分比
日期	0	0%
销售员	0	0%
产品类别	3	15%
产品 ID	1	5%
销售数量	4	20%
单价(元)	3	15%
总金额	6	30%
客户类型	2	10%
地区	3	15%
支付方式	4	20%

(2) 处理步骤。

① 删除缺失值超过 30% 的列：总金额列缺失 30%，达到临界值但未超过，保留该列。

② 数值列填充。

销售数量：使用同月份中位数填充。

单价(元)：使用同月份中位数填充。

③ 分类列填充。

产品类别：缺失值标记为"未知"。

客户类型：缺失值标记为"未知"。

地区：缺失值标记为"未知"。

支付方式：缺失值标记为"未知"。

④ 产品 ID 处理：唯一标识符，删除缺失行(第 11 行)。

(3) 处理后的完整表格。

大模型处理后的完整表格如下表所示。

表 处理后的完整表格

日期	销售员	产品类别	产品 ID	销售数量	单价(元)	总金额	客户类型	地区	支付方式
2023/1/5	张三	电子产品	P1001	3	1200	3600	企业客户	华东	信用卡
2023/1/7	李四	办公用品	P2005	2	45	90	中小企业	华北	现金
2023/1/8	王五	家具	P3002	1	2500	2500	未知	华南	支付宝
2023/1/10	张三	电子产品	P1003	2	800	1600	个人客户	未知	微信
2023/1/12	李四	未知	P2001	5	30	150	中小企业	华东	现金
2023/1/15	王五	家具	P3003	2	1800	3600	企业客户	华北	未知
2023/1/18	赵六	办公用品	P2007	4	60	240	个人客户	华南	信用卡
2023/1/20	张三	未知	P1002	1	1500	1500	企业客户	华东	支付宝
2023/1/22	李四	办公用品	P2003	2	50	100	中小企业	未知	现金
2023/1/28	赵六	电子产品	P1004	1	950	950	个人客户	华南	未知
2023/1/30	张三	办公用品	P2006	2	75	150	企业客户	华东	信用卡
2023/2/2	李四	家具	P3004	2	1950	5000	中小企业	华北	支付宝
2023/2/5	王五	电子产品	P1005	4	1100	4400	未知	华南	微信
2023/2/8	赵六	未知	P2008	3	40	120	个人客户	华东	现金
2023/2/10	张三	办公用品	P2009	1	65	65	企业客户	未知	信用卡
2023/2/12	李四	家具	P3005	2	1950	3900	中小企业	华北	未知
2023/2/15	王五	电子产品	P1006	2	1350	2700	个人客户	华南	支付宝
2023/2/18	赵六	办公用品	P2010	5	60	250	企业客户	华东	微信
2023/2/20	钱七	家具	P3006	1	2800	2800	中小企业	华北	现金

(4) 处理日志。

① 数值列填充详情。

1 月销售数量中位数：2。

2 月销售数量中位数：2。

1 月单价中位数：800(使用办公用品和电子产品的中间值)。

2 月单价中位数：1100。

② 分类列填充详情。

产品类别：填充 3 个"未知"。

客户类型：填充 2 个"未知"。

地区：填充 3 个“未知”。

支付方式：填充 4 个“未知”。

③ 删除记录。

删除了 1 条产品 ID 为空的记录(原第 11 行)。

④ 总金额计算。

对于销售数量和单价都存在的行,重新计算总金额。

其他行保持原值(可能包含促销等特殊定价)。

(5) 处理建议。

① 建议后续收集数据时加强产品类别和支付方式的完整性检查。

② 对于总金额列,虽然保留但需注意 30% 的缺失率可能影响分析结果。

③ 可考虑对支付方式缺失的记录进行进一步调查,可能与特定销售渠道有关。

5.2.2 删除指定的行或列

对提供的表格数据进行行列删除操作的常用提示词如下。

1. 按条件删除行的提示词

按条件删除行的提示词举例如下。

删除表格中[列名]值为[特定值]的所有行。

删除[日期列]早于[2023-01-01]的所有记录。

删除[销售额]低于[1000 元]的行。

删除[客户类型]为空值的所有行。

保留[地区]为['华东','华南']的行,删除其他行。

2. 按位置删除行的提示词

按位置删除行的提示词举例如下。

删除表格的第 5~10 行。

删除首行(标题行除外)和最后两行。

每隔一行删除一行。

删除表格中重复的行,基于[列名]判断重复。

3. 删除列的提示词

删除列的提示词举例如下。

删除[列名 1]、[列名 2]和[列名 3]列。

只保留[列名 A]和[列名 B]列,删除其他所有列。

删除第 3 列和第 5 列。

删除所有包含'备注'字样的列。

删除缺失值超过 30% 的列。

4. 组合操作

组合操作的提示词举例如下。

先删除[行名]为'无效'的行,再删除[列名 1]和[列名 2]两列。

删除第 1~3 行,然后删除[列名]列,最后删除[列名]低于 60 的行。

【例 5-2】 对销售数据表格进行行列处理。

对销售数据表格进行行列处理的提示词如下。

首先删除所有[客户类型]为空的记录。

然后删除[产品 ID]列和[支付方式]列。

接着删除[销售数量]为 0 或负值的行。

最后只保留 2023 年 1 月之后(不含 1 月)的数据。

输出处理后的表格。

大模型处理后的数据表如下。

大模型回复

表 大模型处理后的数据

日期	销售员	产品类别	销售数量	单价(元)	总金额	客户类型	地区
2023-02-02	李四	家具	2	1950	5000	中小企业	华北
2023-02-05	王五	电子产品	4	1100	4400	未知	华南
2023-02-08	赵六	未知	3	40	120	个人客户	华东
2023-02-10	张三	办公用品	1	65	65	企业客户	未知
2023-02-15	王五	电子产品	2	1350	2700	个人客户	华南
2023-02-20	钱七	家具	1	2800	2800	中小企业	华北

5.2.3 重复行处理

对提供的表格数据进行重复行处理的常用提示词如下。

1. 常规去重

表格重复行处理的常规去重的提示词举例如下。

删除表格中完全相同的重复行。

基于[列名]列删除重复行,保留第一次出现的记录。

根据[ID 列]和[日期列]组合判断重复,删除重复行。

保留[客户编号]唯一的记录,删除后续重复项。

2. 高级去重

表格重复行处理的高级去重的提示词举例如下。

删除重复行,但保留[销售额]最高的记录。

当[产品 ID]重复时,保留[更新时间]最新的记录。

对重复的[客户名称]记录,合并其[订单金额]并保留一条记录。

3. 选择性去重

表格重复行处理的选择性去重的提示词举例如下。

仅对 2023 年的数据删除重复行。

在[部门]为'Sales'的记录中删除重复项。

删除[状态]为'已完成'的重复订单记录。

【例 5-3】 表格去重举例。

销售数据如表 5-2 所示。

表 5-2 销售数据

订单编号	客户名称	产品名称	订单金额	下单日期	销售员
ORD1001	张伟	笔记本电脑	5299	2023/3/10	王丽
ORD1002	李娜	智能手机	3899	2023/3/11	张强
ORD1003	王芳	平板计算机	2599	2023/3/12	王丽
ORD1001	张伟	笔记本电脑	5299	2023/3/10	王丽
ORD1004	赵刚	智能手表	1299	2023/3/13	李婷
ORD1002	李娜	智能手机	4199	2023/3/14	张强
ORD1005	孙丽	耳机	599	2023/3/15	王丽
ORD1003	王芳	平板计算机	2599	2023/3/12	王丽

对表 5-2 的销售数据进行去重处理的提示词如下。

处理提供的表格。

首先基于[订单编号]列识别重复行。

对于重复订单,保留[创建时间]最早的记录。

如果创建时间相同,则保留[订单金额]较大的记录。

记录被删除的重复行数和保留策略。

输出处理后的表格及去重报告。

大模型处理后的表格如表 5-3 所示。

表 5-3 大模型处理后的表格

订单编号	客户名称	产品名称	订单金额	下单日期	销售员	保留原因
ORD1001	张伟	笔记本电脑	5299	2023/3/10	王丽	首次出现
ORD1002	李娜	智能手机	3899	2023/3/11	张强	较早日期
ORD1003	王芳	平板计算机	2599	2023/3/12	王丽	首次出现
ORD1004	赵刚	智能手表	1299	2023/3/13	李婷	唯一记录
ORD1005	孙丽	耳机	599	2023/3/15	王丽	唯一记录

5.2.4 数据变换

1. 常用数据变换类型及提示词

(1) 数值计算与转换提示词。

数值计算与转换提示词举例如下。

新增[利润率]列,计算公式为(销售额-成本)/销售额。

将[价格]列从美元转换为人民币(汇率 6.8)。

对[年龄]列进行分箱处理: 0~18 为'未成年',19~35 为'青年',36~55 为'中年',56+为'老年'。

(2) 文本处理提示词。

文本处理提示词举例如下。

将[姓名]列拆分为[姓氏]和[名字]两列。

将[地址]列中的'市' '区'等字样统一删除。

将[产品描述]列全部转换为小写形式。

(3) 日期时间处理提示词。

日期时间处理提示词举例如下。

从[日期时间]列中提取出[日期]和[时间]两列。

计算[下单日期]与[发货日期]之间的天数差,存入[配送时长]列。

将[季度]列(如 Q1)转换为实际日期范围(如 2023-01-01 至 2023-03-31)。

(4) 分类编码提示词。

分类编码提示词举例如下。

将[产品类别]列进行 One-Hot 编码。

对[满意度评分]列(1~5)转换为分类标签: 1~2 为'不满意', 3 为'一般', 4~5 为'满意'。

将[地区]列中的中文省份名称转换为对应的省份编码。

(5) 行列转换提示词。

行列转换提示词举例如下。

将表格从长格式转为宽格式,以[产品 ID]为行,[月份]为列。

转置表格,使原列名成为第一列数据。

将多列[1 月销售额]、[2 月销售额]、……、[12 月销售额]合并为[月份]和[销售额]两列的长格式。

2. 高级变换提示词

(1) 条件变换提示词。

条件变换提示词举例如下。

当[销售额]>10000 时,[折扣率]列设为 15%,否则设为 10%。

如果[库存量]<10,则[状态]列标记为'需补货',否则保持原值。

(2) 多列联动变换提示词。

多列联动变换提示词举例如下。

根据[身高]和[体重]列计算 BMI 值,并新增[健康状态]列(BMI<18.5 为'偏瘦',18.5~24 为'正常',>24 为'超重')。

将[城市]列和[区域]列合并为新列[完整地址],格式为'城市-区域'。

(3) 自定义函数应用提示词。

自定义函数应用提示词举例如下。

对[评论内容]列应用情感分析,新增[情感倾向]列(正面/中性/负面)。

使用以下规则标准化[电话号码]列: 去除所有非数字字符,前加国家代码 86。

【例 5-4】 数据变换举例。

销售数据如表 5-4 所示。

对表 5-4 的销售数据进行数据变换的提示词如下。

请对销售数据表进行以下变换处理。

(1) 将[订单日期]拆分为[年]、[月]、[日]三列。

表 5-4 销售数据

订单 ID	客户名称	产品类别	单价	数量	订单日期	客户类型	地区
1001	张伟	电子产品	1200	2	2023/3/10	企业客户	华东
1002	李娜	办公用品	45	5	2023/3/11	个人客户	华北
1003	王芳	家具	2500	1	2023/3/12	中小企业	华南
1004	赵刚	电子产品	800	3	2023/3/13	企业客户	华东
1005	孙丽	办公用品	60	4	2023/3/14	个人客户	华北

(2) 新增[销售额区间]列。

<1000: '低'

1000~5000: '中'

5000: '高'

(3) 计算[利润率]列: (销售额-成本)/销售额,保留 2 位小数。

(4) 对[客户类型]列进行标签编码。

'企业客户'→1

'个人客户'→2

'中小企业'→3

(5) 将处理后的数据保存为新表格。

大模型处理后的表格如表 5-5 所示。

表 5-5 大模型处理后的表格

订单 ID	客户名称	产品类别	单价	数量	年	月	日	销售额	销售额区间	利润率	客户类型编码	地区
1001	张伟	电子产品	1200	2	2023	3	10	2400	中	0.2	1	华东
1002	李娜	办公用品	45	5	2023	3	11	225	低	0.2	2	华北
1003	王芳	家具	2500	1	2023	3	12	2500	中	0.2	3	华南
1004	赵刚	电子产品	800	3	2023	3	13	2400	中	0.2	1	华东
1005	孙丽	办公用品	60	4	2023	3	14	240	低	0.2	2	华北

5.3 统计分析

设计统计分析提示词时,需明确分析目标、数据背景和所需结果类型,充分利用大模型的自然语言理解能力,用结构化描述表达提高大模型回答的准确性。统计分析提示词设计的主要原则如下。

1. 明确统计方法的类型

在设计统计分析提示词时,清晰指定统计方法类型至关重要,如描述性统计、相关性分析、回归模型等,这能帮助大模型精准理解需求并返回正确结果。

(1) 描述性统计：用于概括数据的基本特征,适用于初步了解数据分布情况,常用指标有均值、中位数、众数、标准差、极差、四分位数、偏度、峰度等。描述性统计主要用于数据探索、报告摘要撰写场景。

【例 5-5】 对表 5-6 所示的小麦产量进行描述性统计的提示词：计算小麦产量的均值、中位数、标准差和四分位数,并输出表格。

表 5-6 小麦产量

指 标	小麦产量(万吨)	指 标	小麦产量(万吨)
1990 年	9822.9	2007 年	10949.15
1991 年	9595.3	2008 年	11290.13
1992 年	10158.7	2009 年	11579.61
1993 年	10639	2010 年	11609.34
1994 年	9929.7	2011 年	11856.95
1995 年	10220.7	2012 年	12247.49
1996 年	11056.91	2013 年	12363.93
1997 年	12328.85	2014 年	12823.52
1998 年	10972.61	2015 年	13255.52
1999 年	11387.96	2016 年	13318.83
2000 年	9963.58	2017 年	13424.13
2001 年	9387.34	2018 年	13144.05
2002 年	9029	2019 年	13359.63
2003 年	8648.85	2020 年	13425.38
2004 年	9195.18	2021 年	13694.45
2005 年	9744.51	2022 年	13772.34
2006 年	10846.59	2023 年	13659

大模型回复的表格如表 5-7 所示。

表 5-7 大模型回复的表格

统 计 量	值(万吨)	说 明
数据量(年份)	34	1990—2023 年完整数据
均值	11302.98	34 年平均年产量
中位数	11044.26	数据中间值(50%分位数)
标准差	1547.50	产量波动程度
最小值	8648.85	2003 年最低产量
25%分位数	9744.51	2005 年数据(第一四分位数)
75%分位数	12328.85	1997 年数据(第三四分位数)
最大值	13772.34	2022 年最高产量

(2) 相关性分析：用于衡量变量间的关联程度,如研究广告投入与销售额的关系、用户