项目三
商务数据预处理

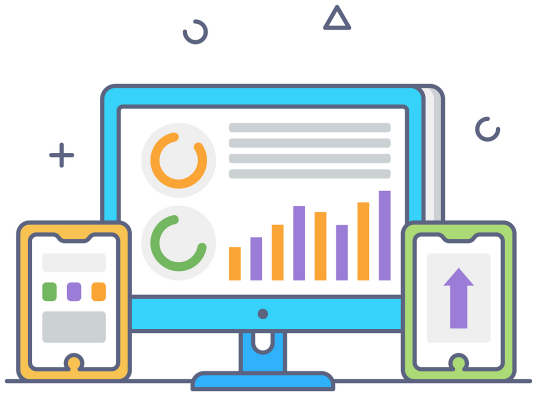
了解商务数据处理的概念；

学会查找原数据存在的主要问题；

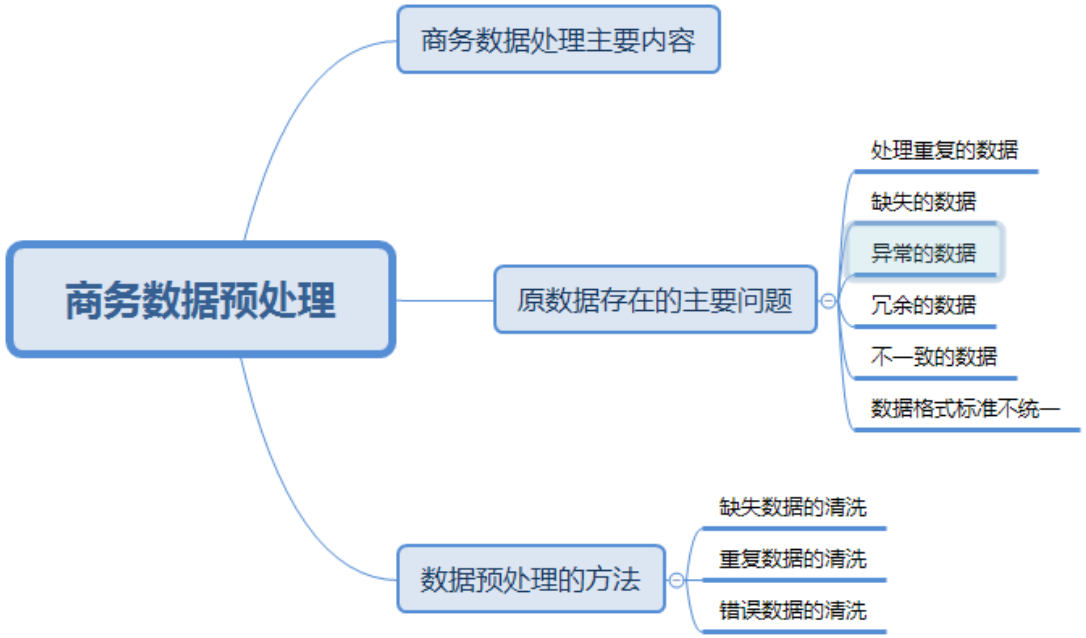
掌握数据预处理的方法。

项目介绍

项目思维导图



Data Analysis



【案例导入】

我国至少有1.5亿只宠物，宠物经济的市场潜力也达到了250亿人民币，宠物用品蕴藏着巨大的商机。梅青也趁热在淘宝平台开了一家宠物用品店，为了制订更好的营销策略，她每个月都要对新增的店内数据与历史数据进行汇总。如何快速地完成这个重复工作呢？如何运用EXCEL工具，对店铺运营中大量的数据进行提取有用的数据呢？

数据处理有广义和狭义之分。广义的数据处理包括所有的数据采集、存储、加工、分析、挖掘和展示等工作；而狭义的数据处理仅仅包括从存储的数据中提取筛选出有用的数据，对有用的数据进行加工的过程。一般数据处理就是对数据进行增加、删除、修改、查询等操作。

目 录

CONTENTS

3.1 任务一 了解商务数据处理主要内容

任务描述

张雷根据指导专家的建议，在部门同事的协调合作下，制定了明确的商业规划过程，但制定规划只是目标实施的第一步，接下来，张雷面临更大的挑战，即是如何获取所需的特定数据，获取数据之后又该对数据进行哪些处理呢，数据处理过程中需要注意哪些问题，运用哪些方法，带着这些问题，张雷重新开始了“取经”之路。

广义的数据处理

数据处理有广义和狭义之分，广义的数据处理包括所有的数据采集、存储、加工、分析、挖掘和展示等工作，而狭义的数据处理仅仅包括从存储的数据中提取筛选出有用的数据，对有用的数据进行加工的过程。

狭义的数据处理

常用的数据处理是狭义的定义，即对数据进行增加、删除、修改、查询等操作。 在目前大量数据背景下，数据处理工作往往是通过技术手段来实现的，如利用数据库对处理的数据进行增加、删除、修改、查询等。

在数据处理过程中最大的工作是对数据进行清洗，即将不清洁的数据进行清洁化，使得数据理更加规范，让数据结构更加合理，让数据的含义更加明确，并让数据处理于数学模型的可用状态，

狭义的数据处理

通过各种渠道收集来的数据，常出现缺失、异常、冗余、不一致等现象，质量并不能直接为数据分析所用。此外，一些成熟的数据分析模型对处理的数据有特定的要求，比如一定的数据类型、统一的数据量纲、数据的冗余性要求、属性的相关性要求等。

原始数据存在的问题包括： 重复数据、缺失数据、异常值、冗余数据、不一致数据。

目 录

CONTENTS

3.2.1 重复数据

重复数据是指在数据表中唯一标识记录的字段出现多次的数据。

例如，在图 3-1 所示的会员信息表中，会员编号是可以唯一标识每条记录的指标。其中，会员编号“1893133”出现了 2 次，为重复数据。

会员信息表

会员编号	年龄	性别	联系手机	收货地址
97485	40	男	137****8004	广东省 深圳市 大鹏新区
190695	45	女	158****5099	广东省 深圳市 福田区
489376	30	女	136****8028	广东省 深圳市 龙岗区
1893133	29	男	137****0703	广东省 广州市 白云区
493834	47	女	139****2634	广东省 深圳市 龙岗区
558903	36	female	187****4577	广东省 中山市 小榄镇
559569	48	male	135****1048	广东省 广州市 花都区
893869	29	女	186****1665	广东省 梅州市 梅县区
1333727	33	female	181****8906	广东省 广州市 白云区
1893133	29	男	137****0703	广东省 广州市 白云区
2263904	33	男	150****1945	广东省 深圳市 罗湖区
2310007	32	女	186****0221	广东省 深圳市 南山区
2490531	33			
2689842	30	男	151****5892	广东省 深圳市 龙岗区
2925852	27	男	138****0278	广东省 中山市 小榄镇
3061820	20	女	138****6726	广东省 阳江市 江城区
3139245	24	男	151****8770	广东省 东莞市 东城街
3149821	42	女	138****6726	广东省 深圳市 福田区
4153242	32	女	137****2026	广东省 东莞市 石龙镇
4153485	34	女	156****8632	广东省 梅州市 五华县
4153595	24	男	150****9225	广东省 江门市 江海区
4290542	30	女	134****8287	广东省 深圳市 龙岗区
4313145	33	女	183****1965	广东省 东莞市 石龙镇
4313496	26	女	151****8770	广东省 东莞市 东城街
4372630	41	male	159****9830	广东省 广州市 黄埔区
4414023	40	男	135****1029	广东省 梅州市 五华县
4717133	43	女	134****0987	广东省 惠州市 惠阳区
4741189	34	男	137****5519	广东省 惠州市 博罗县
4855701	23	女	187****8120	广东省 惠州市 惠阳区
4936166	23	女	137****2803	广东省 湛江市 赤坎区
5515369	44	female	137****5735	广东省 惠州市 惠阳区

3.2.2 缺失数据

缺失数据是指在实践过程中因种种原因没有能够获取观测对象的相关信息，造成数据不完全。

例如数据录入、存储过程中的人为失误和系统软硬件问题造成了数据的缺失等。

会员信息表

会员编号	年龄	性别	联系手机	收货地址
97485	40	男	137****8004	广东省 深圳市 大鹏新区
190695	45	女	158****5099	广东省 深圳市 福田区
489376	30	女	136****8028	广东省 深圳市 龙岗区
1893133	29	男	137****0703	广东省 广州市 白云区
493834	47	女	139****2634	广东省 深圳市 龙岗区
558903	36	female	187****4577	广东省 中山市 小榄镇
559569	48	male	135****1048	广东省 广州市 花都区
893869	29	女	186****1665	广东省 梅州市 梅县区
1333727	33	female	181****8906	广东省 广州市 白云区
1893133	29	男	137****0703	广东省 广州市 白云区
2263904	33	男	150****1945	广东省 深圳市 罗湖区
2310007	32	女	186****0221	广东省 深圳市 南山区
2490531	33			
2689842	30	男	151****5892	广东省 深圳市 龙岗区
2925852	27	男	138****0278	广东省 中山市 小榄镇
3061820	20	女	138****6726	广东省 阳江市 江城区
3139245	24	男	151****8770	广东省 东莞市 东城街
3149821	42	女	138****6726	广东省 深圳市 福田区
4153242	32	女	137****2026	广东省 东莞市 石龙镇
4153485	34	女	156****8632	广东省 梅州市 五华县
4153595	24	男	150****9225	广东省 江门市 江海区
4290542	30	女	134****8287	广东省 深圳市 龙岗区
4313145	33	女	183****1965	广东省 东莞市 石龙镇
4313496	26	女	151****8770	广东省 东莞市 东城街
4372630	41	male	159****9830	广东省 广州市 黄埔区
4414023	40	男	135****1029	广东省 梅州市 五华县
4717133	43	女	134****0987	广东省 惠州市 惠阳区
4741189	34	男	137****5519	广东省 惠州市 博罗县
4855701	23	女	187****8120	广东省 惠州市 惠阳区
4936166	23	女	137****2803	广东省 湛江市 赤坎区
5515369	44	female	137****5735	广东省 惠州市 惠阳区

缺失数据

3.2.3 异常值

异常值也可称为离群点，是指所获得的数据中与平均值的偏差超过两倍及两倍以上标准差的数据。

异常值产生的原因很多，例如录入数据时误将“80”录入为“800”，那么当数据均为100 左右的数据时，“800”就会被识别为异常值。

当异常值存在时，会严重影响数据分析的结果，例如使平均值偏高或偏低，使方差增大，影响数据模型的拟合优度等。此外，若异常值不是错误数据，就应是数据分析人员关注的焦点。

3.2.4 冗余数据

数据冗余一方面是指多个数据集合并时同一条数据命名或者编码方式不同，例如某数据集中的变量名称为“用户编码”而在另一个数据集中为“id”；另一方面指数据集中的2 个或多个变量之间存在相关或者推导关系。冗余数据会造成数据重复或分析结果产生偏差。

3.2.5 不一致数据

一是人为 / 机械原因导致的录入错误或数据规范不同，例如将数据集中的“客单价”录入为“-180”；又如变量名“用户编码”下，某数据集的规范是“3 位 / 数字”，在另一个数据集中则要求“5 位 / 字母 + 数字”。

二是变量单位或者量纲不匹配。例如，某数据集中的商品价格以“元”为单位，另一个数据集中却为“万元”。

三是数据特征不适应特定数据分析模型的需求或变量过多，分析难度较大。例如，手机系统分为 Android 和 iOS 两种，但回归分析模型中要求数据是数值型的，可以将其转换为名义变量（0/1 变量）再进行处理。

3.2.6 数据格式标准不统一

所谓数据格式标准不统一是指在录入数据时使用了错误的格式。 例如在录入日期时，因为格式不规范，计算机不能自动识别为日期格式。 这样的情况常处理方式为，在信息系统中设定相关的数据校验如果录入的数据格式不一致时则系统会弹出数据录入格式错误的警告。

目 录

CONTENTS

3.3 任务三 掌握数据预处理的方法

针对原始数据存在的问题，我们需要执行商务数据分析工作流程中一个必不可少的环节——商务数据预处理。商务数据预处理的意义主要有以下两个方面。

（1）挖掘商务数据特征，提高原始数据的质量。

（2）为后续的商务数据分析提供必要的数据形式。

预处理方法	关注的数据问题
统计特征处理	集中趋势、离散趋势、异常值
商务数据清洗	重复数据、缺失数据、异常值
商务数据集成、转换和规约	冗余数据、不一致数据

3.3 任务三 掌握数据预处理的方法

（1）商务数据统计特征处理

是指对数据总体或者对感兴趣的目标总体的集中趋势和离散程度进行测度，从整体上把握总体的基本特征、相关指标（如总体均值、总体方差、总体变异系数）等。

（2）商务数据清洗

是指对数据集中可能存在的重复数据、缺失数据及异常值进行必要的处理。

（3）商务数据集成、转换和规约

商务数据集成也可称为数据整合，是对同一目标总体不同来源、异构的数据的合并。

商务数据转换是指将数据转换成统一的、适用于数据分析方法应用的数据形式。

商务数据规约是指在尽量保证原数据完整性的前提下将数据集的规模缩小，以提高数据分析的效率。

3.3.1重复数据的检测与删除

重复数据的检测方法有很多，以 Excel 为例，可以采用筛选、条件格式设置及 COUNTIF 函数实现。删除重复数据则可采用“删除重复项”功能、排序或筛选方式完成。

他山之石：

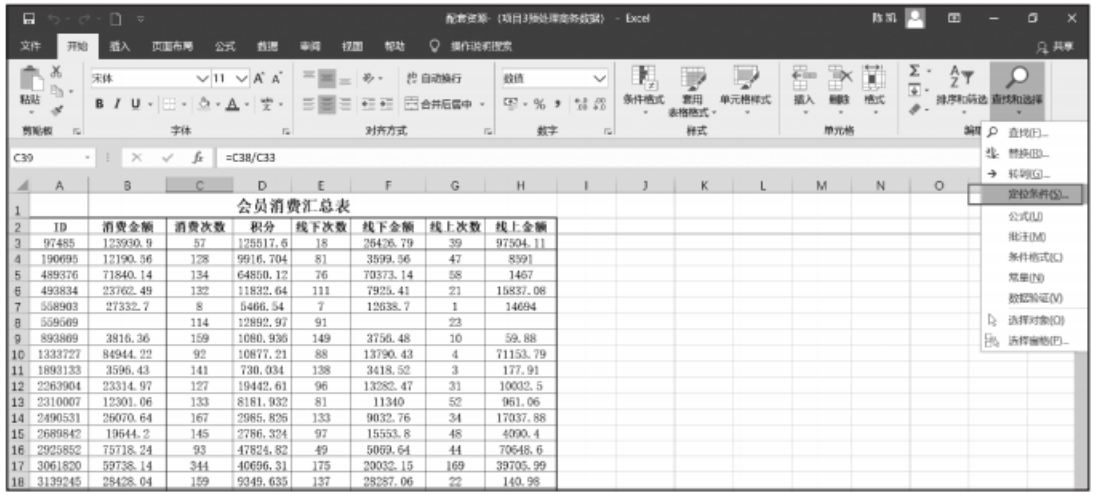
（1）利用 Excel 对重复数据进行检测

（2）利用 Excel 对重复数据进行删除

3.3.2缺失数据的检测与处理

缺失数据一般在数据表中表现为空白单元格或错误标识符。

其中，空白值在Excel 软件中可单击“开始”选项卡的“编辑”功能区，通过“定位”→“定位条件”→“空值”→“确定”，将缺失数据一次性选定。错误标识符则需根据存储文件特征查找原因，例如 Excel 中，“####”表示单元格中的数据超出了该单元格的宽度，或者单元格中的日期时间公式产生了负值 ；“#DIV/0！”表示进行公式运算时，除数使用了数值零、指向了空单元格等。



3.3.2缺失数据的检测与处理

缺失数据的处理主要有 4 种方法

① 用样本统计量代替缺失数据，最典型的做法是使用变量的平均值替代。替代后由于该变量的平均值会保持不变，因此其他的统计量（如标准差和相关系数等）也不会受很大的影响。

② 将有缺失数据的记录删除，删除记录会导致样本量减少，所以此方法不适于小样本量的数据集。

③ 将有缺失数据的记录保留，仅在相应的分析中做必要的排除。当调查的样本量比较大，缺失数据的数量又不是很多，而且变量之间也不存在高度相关的情况下，采用这种方式处理缺失数据比较可行。

④ 利用由某些统计模型计算得到的比较合理的值来代替，例如利用回归模型、判别分析模型等。

3.3.2缺失数据的检测与处理

想一想：

一组数据为　　3、31、15、9、17、24、8、28、（　）。假设（　）中的值是缺失值，那么该如何处理?

3.3.2缺失数据的检测与处理

他山之石：

（1）利用 Excel 对缺失数据进行检测

（2）利用 Excel 对缺失数据进行处理

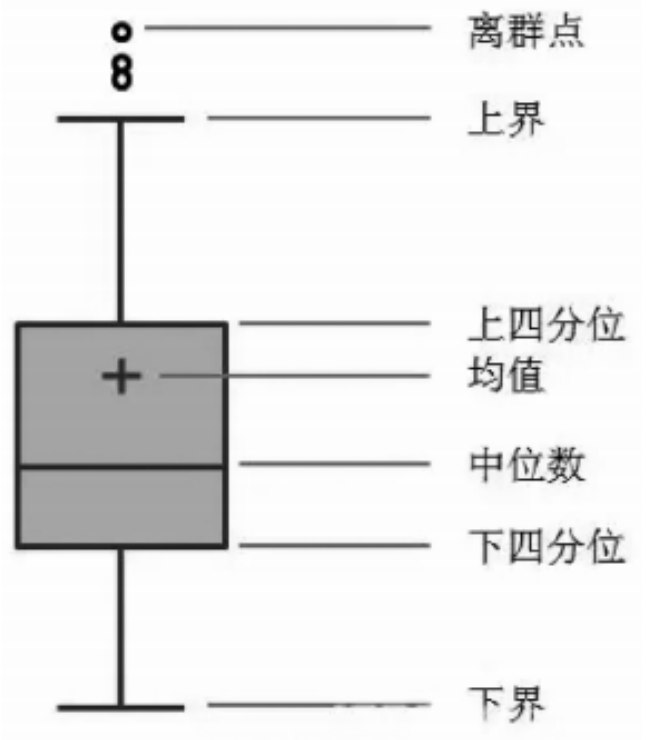
3.3.3异常值的检测与处理

异常值可通过数据的统计特征初步识别，一般偏离数据集的平均值较大的即为异常值；如能将数据集可视化，也可以从图表中直观地发现异常值。

异常值检测在数据分析过程中有重要的意义，如能回溯确认数据是人工 / 机械录入错误则可直接修正为真实值；又如异常值是由于数据本身的变异造成的，那么对其进行分析，就可以发现隐藏的更深层次的、潜在有价值的信息。

3.3.3异常值的检测与处理

箱线图是由数据的上边缘、上四分位数、中位数、下四分位数和下边缘组成的图形，其中上边缘和下边缘线所代表的就是临界值，超过上下边界的离群点则为需要关注的异常值。



3.3.3异常值的检测与处理

对异常值在处理时可采用以下方法

① 参考后续的数据分析模型，选择删除或者保留异常值。

② 用一个样本统计量去代替异常值，比如平均值、中位数、众数等。

③ 分箱法，即通过考察相邻数据的取值对异常值进行平滑处理，可视为一种局部平滑方法。首先将异常值所在指标下的所有数据按照大小排序，并适当分组（也称作分“箱”），然后用组内数据的平均值、中位数或边界值来代替异常值。分组时，如果每个“箱”的数据个数相同就为等深分箱；如果每个“箱”内数据值的区间范围是一个常量就为等宽分箱。

3.3.3异常值的检测与处理

例如，设定箱深为 3，对数据集 {150、100、80、200、180、280、450、500、350} 进行等深分箱，结果如下。

箱 1：{80、100、150}

箱 2：{180、200、280}

箱 3：{350、450、500}

设定区间范围为 100，对数据集 {150、100、80、200、180、280、450、500、350} 进行等宽分箱，结果如下。

箱 1：{80、100、150、180}

箱 2：{200、280}

箱 3：{350、450}

箱 4：{500}

3.3.3异常值的检测与处理

想一想：

一组数据如下：3、31、15、9、17、24、8、28、105。假设 105 是异常值，该如何处理?

3.3.3异常值的检测与处理

他山之石：

（1）利用 Excel 对异常数据进行检测

（2）使用分箱法在 Excel 中对异常数据进行处理

Thank You!