**Pandas预处理任务要求：**

**1.堆叠不同时间的订单详情表**

**订单详情表meal\_order\_detail1、meal\_order\_detail2、meal\_order\_detail3具有相同的特征，但数据时间不同，订单编号也不同，在数据分析过程中需要使用全量数据，故需要将几张表做纵向堆叠操作。**

**2.对订单详情表进行去重**

**菜品数据存在重复的现象，所以需要对菜品销售数据分别进行记录去重和特征去重（**通过DataFrame.equals的方法**）。**

**3.处理订单详情表缺失值**

**统计经过去重操作后的各个特征缺失值，然后根据缺失值情况选择缺失值处理方法。**

**（要求：删除全部数据均为缺失值的列）**

**4.处理菜品销售数据异常值**

**通过探索知道，在counts, amount特征中存在异常值(用箱线图法进行识别)。使用替换法处理这些异常值。**

**（要求：用QL和QU替换异常值），具体参考课件和课上源程序。**

**5.主键合并订单详情表、订单信息表和客户信息表**

**订单详情表、订单信息表和客户信息表两两之间存在相同意义的字段，因此需通过主键合并的方式将三张表合并为一张宽表（合并后主键只保留一个）。**

**订单详情表：order\_id（订单ID）**,**emp\_id（会员ID）**

**订单信息表：info\_id（订单ID）**,**emp\_id（会员ID）**

**客户信息表： USER\_ID（用户ID）**

**注：会员ID和用户ID是相同含义。**

**输出结果参考如下：**

三张订单详情表合并后的形状为： (10037, 19)

进行去重操作前订单详情表的形状为： (10037, 19)

样本去重后订单详情表的形状为： (10037, 19)

通过DataFrame.equals的方法得到特征矩阵并保存到文件detEquals\_js.xlsx中.

根据特征矩阵确定需要删除的列有： ['parent\_class\_name', 'cost', 'discount\_amt', 'discount\_reason', 'kick\_back', 'add\_info', 'bar\_code', 'add\_inprice']

进行去重操作后订单详情表的形状为： (10037, 11)

detail每个特征缺失值数量为：

detail\_id 0

order\_id 0

dishes\_id 0

logicprn\_name 10037

dishes\_name 0

itemis\_add 0

counts 0

amounts 0

place\_order\_time 0

picture\_file 0

emp\_id 0

dtype: object

删除全部数据均为缺失的列后订单详情表的形状为： (10037, 10)

异常值处理前销售量最小值为： 1

异常值处理前销售量最大值为： 10

异常值处理前售价最小值为： 1

异常值处理前售价最大值为： 178

异常值处理后销售量最小值为： 1.0

异常值处理后销售量最大值为： 1.0

异常值处理后售价最小值为： 1.0

异常值处理后售价最大值为： 99.0

detail表大小： (10037, 10)

order表大小： (945, 21)

user表大小： (758, 37)

三张表数据主键合并后的大小为： (14713, 65)

合并后的宽表的列： Index(['detail\_id', 'order\_id', 'dishes\_id', 'dishes\_name', 'itemis\_add',

'counts', 'amounts', 'place\_order\_time', 'picture\_file', 'emp\_id',

'number\_consumers', 'mode', 'dining\_table\_id', 'dining\_table\_name',

'expenditure', 'dishes\_count', 'accounts\_payable', 'use\_start\_time',

'check\_closed', 'lock\_time', 'cashier\_id', 'pc\_id', 'order\_number',

'org\_id', 'print\_doc\_bill\_num', 'lock\_table\_info', 'order\_status',

'phone', 'name', 'MYID', 'ACCOUNT', 'NAME', 'ORGANIZE\_ID',

'ORGANIZE\_NAME', 'DUTY\_ID', 'TITLE\_ID', 'PASSWORD', 'EMAIL', 'LANG',

'THEME', 'FIRST\_VISIT', 'PREVIOUS\_VISIT', 'LAST\_VISITS', 'LOGIN\_COUNT',

'ISEMPLOYEE', 'STATUS', 'IP', 'DESCRIPTION', 'QUESTION\_ID', 'ANSWER',

'ISONLINE', 'CREATED', 'LASTMOD', 'CREATER', 'MODIFYER', 'TEL', 'stuNo',

'qq', 'weixin', 'meal\_arithmetic\_id', 'arithmetic\_name', 'sex', 'poo',

'address', 'age'],

dtype='object')

