# 数据加载、存储与文件格式



索引：将一个或多个列当做返回的Dataframe处理，以及是否从文件、用户获取列名；

类型推断和数据转换：包括用户定义值得转换、缺失值标记列表等；

日期解析：包括组合功能，比如将分散在多个列中的日期时间信息组合成结果中的单个列；

迭代：支持对大文件进行逐块迭代；

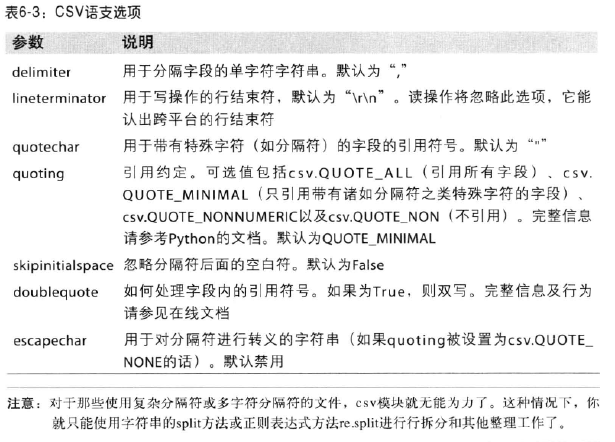
不规整数据问题：跳过一些行、页脚、注释或其他一些不重要的东西。

1. 定义”表头”信息
2. 定义“索引”信息
3. 多个列作为一个层次化索引
4. 使用正值表达式作为字段间分隔符





* 逐块读取文本文件
* 将数据写出到文本格式
* 手工处理分隔符格式



* JSON数据
* XML和HTML：web信息收集

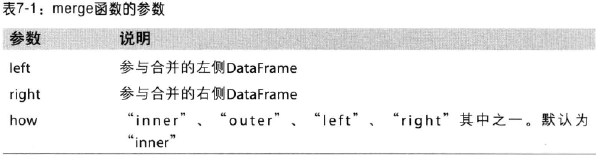
lxml.html处理Html；lxml.objectify处理XML

* 利用lxml.objectify解析XML
* 使用HDF5格式
* 读取Microsoft Excel文件
* 使用数据库
* 存取MongoDB中的数据

# 数据规整化：清理、转换、合并、重塑

* 合并数据集
* 数据库风格的Dataframe合并

数据集的合并（merge）或连接（join）运算时通过一个或多个键将行链接起来。



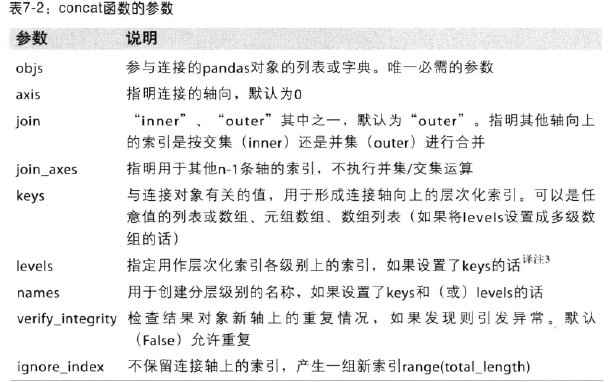


* 索引上的合并

Dataframe中的链接位与其索引中，通过传入left\_index=true或right\_index=true，说明引用该被用作链接链。

* 轴向链接

另一种数据合并运算也被称作连接（concatenation）、绑定（binding）或堆叠（stacking）



* 合并重叠数据

可能有索引全部或部分重叠的两个数据集。

* 重塑和轴向旋转-用于重新排列表格型数据。
* 重塑层次化索引

Stack：将数据的列“旋转”为行；

UnStack：将数据的行“旋转”为列；

* 将“长格式”旋转为“宽格式”

时间序列数据通常称为“长格式”或“堆叠格式”存储在数据和csv中。

DataFrame的pivot方法实现

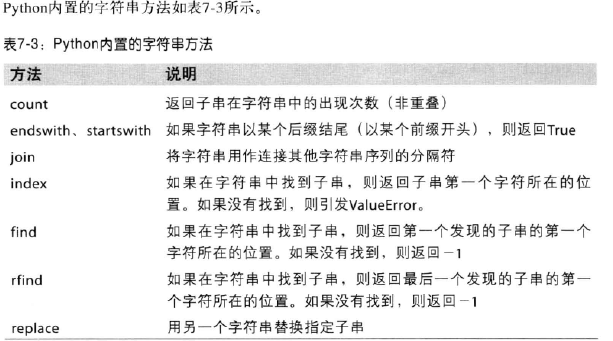
* 数据转换-过滤、清理以及其他的转换工作
* 移除重复数据：duplicated
* 利用函数或映射进行数据转换：map实现元素级转换以及其他数据清理工作。
* 替换值：fillna方法填充缺失数据；map用于修改对象的数据子集；

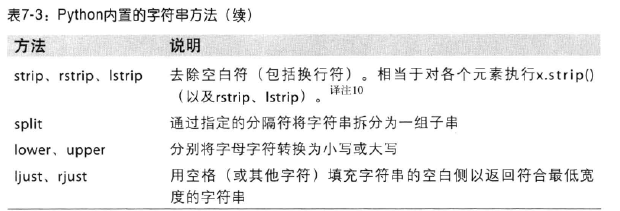
Replace实现以上两函数功能

* 重命名索引：rename
* 离散化和面元划分：为了便于分析，连续数据将被离散化或拆分为“面元”。

Pandas的cut，返回categorical对象类型。

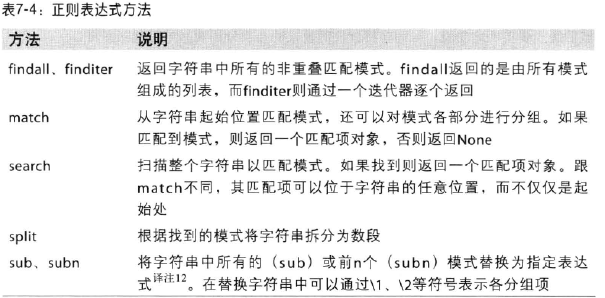
* 检测和过滤异常值
* 排列和随机采样：numpy.random.permutaion
* 计算指标、哑变量：将分类变量（categorical variable）转换为“哑变量矩阵”或“指标矩阵”
* 字符串操作



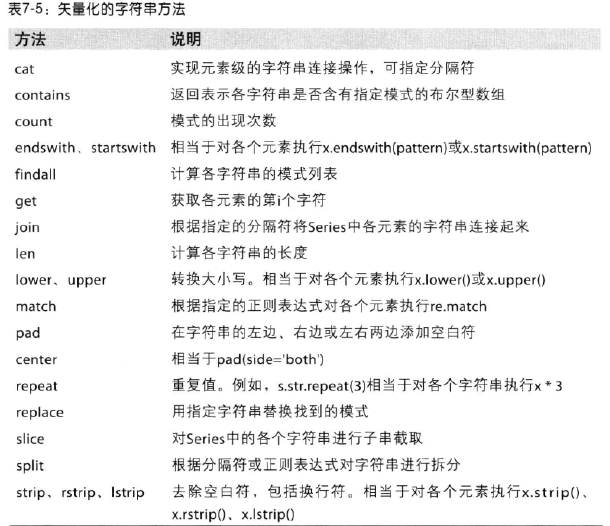


* 正则表达式

Python内置的re模块负责对字符串应用正则表达式。Re模块的函数分为：模式匹配、替换以及拆分。



* Pandas中矢量化的字符串函数



# 绘图和可视化

# 数据聚合与分组运算

目标：

1. 根据一个或多个键（可以是函数、数组或DataFrame列名）拆分对象；
2. 计算分组摘要统计，如计数、平均值、标准差、或用户自定义函数；
3. 对DataFrame的列应用各种各样的函数；
4. 应用组内转换或其他运算，如规格化、线性回归、排名或选取子集等；
5. 计算透视表或交叉表；
6. 执行分位数以及其它分组分析。

# 时间序列

# 金融和经济数据应用