Python for Data Analysis 3rd

Stephen CUI

January 7, 2024

Contents

Chapter 1

数据聚合与分组运算

1.1 Data Aggregation

1.2 apply：一般性的“拆分-应用-合并”

如果传给 apply 的函数能够接受其他参数或关键字，则可以将这些内容放在函数名后面一并传入。传入的那个函数能做什么全由你说了算，它只需返回一个pandas对象或标量值即可。

1.2.1 禁止分组键

1.2.2 分位数和封箱分析

由 pd.cut（equal-length）返回的 pd.Categorical 对象可直接传递到 groupby。

Table 1.1: Optimized groupby methods

Function name	Description
any, all	Return True if any (one or more values) or all non-NA values are “truthy”
count	Number of non-NA values
cummin, cummax	Cumulative minimum and maximum of non-NA values
cumsum	Cumulative sum of non-NA values
cumprod	Cumulative product of non-NA values
first, last	First and last non-NA values
mean	Mean of non-NA values
median	Arithmetic median of non-NA values
min, max	Minimum and maximum of non-NA values
nth	Retrieve value that would appear at position n with the data in sorted order
ohlc	Compute four “open-high-low-close” statistics for time series-like data
prod	Product of non-NA values
quantile	Compute sample quantile
rank	Ordinal ranks of non-NA values, like calling Series.rank
size	Compute group sizes, returning result as a Series
sum	Sum of non-NA values
std, var	Sample standard deviation and variance

Chapter 2

Time Series

2.1 日期和时间数据类型和工具

2.1.1 字符串和日期时间之间的转换

你可以使用或方法将 datetime 对象和 pandas Timestamp 对象格式化为字符串，并传递格式规范：str 或 strftime。

datetime对象还具有许多针对其他国家或语言的系统的特定于区域设置的格式选项。例如，与英语系统相比，德语或法语系统上的缩写月份名称会有所不同。

2.2 时间序列基础知识

pandas 中的一种基本时间序列对象是按时间戳索引的 Series。

pandas.Timestamp 大多数需要使用对象的地方都可以用来代替 datetime。然而，反之则不然，因为 pandas.Timestamp 可以存储纳秒精度的数据，而 datetime 只能存储最多微秒的数据。

和以前一样，你可以传递字符串日期、日期 datetime 或时间戳。请记住，以这种方式切片会生成源时间序列的视图，就像切片 NumPy 数组一样。这意味着不会复制任何数据，并且对切片的修改将反映在原始数据中。

There is an equivalent instance method, `truncate`, that slices a Series between two dates。

这些操作对 DataFrame 同样适用。

Table 2.1: datetime格式规范（ISO C89 兼容）

类型	描述
%Y	四位数年份
%y	两位数年份
%m	两位数月份 [01, 12]
%d	两位数的日期 [01, 31]
%H	小时（24 小时制）[00, 23]
%I	小时（12 小时制）[01, 12]
%M	两位数分钟 [00, 59]
%S	Second [00, 61]（秒 60, 61 代表闰秒）
%f	微秒为整数，用零填充（从 000000 到 999999）
%j	一年中的第几天，以零填充的整数（从 001 到 336）
%w	整数形式的工作日 [0（星期日）, 6]
%u	工作日为从 1 开始的整数，其中 1 表示星期一
%U	一年中的周数 [00, 53]；星期日被视为一周的第一天，一年中第一个星期日之前的日子是“第 0 周”
%W	一年中的周数 [00, 53]；星期一被视为一周的第一天，一年中第一个星期一之前的日子是“第 0 周”
%z	UTC 时区偏移量为+HHMM或-HHMM；如果时区天真则为空
%Z	时区名称为字符串，如果没有时区则为空字符串
%F	%Y-%m-%d（例如，2012-4-18）的快捷方式
%D	%m/%d/%y（例如，04/18/12）的快捷方式

Table 2.2: 特定于区域设置的日期格式

类型	描述
%a	工作日缩写名称
%A	完整工作日名称
%b	月份名称缩写
%B	完整的月份名称
%c	完整日期和时间（例如“2012 年 5 月 1 日星期二 04:20:57 PM”）
%p	AM 或 PM 的区域设置等效项
%x	适合区域设置的格式化日期（例如，在美国，2012 年 5 月 1 日生成“05/01/2012”）
%X	适合区域设置的时间（例如“04:24:12 PM”）

2.3 日期范围、频率和变化

2.3.1 生成日期范围

2.3.2 频率和日期偏移

pandas中的频率是由一个**基础频率**（base frequency）和一个**乘数**组成的。基础频率通常以一个字符串别名表示，比如"M"表示每月，"H"表示每小时。对于每个基础频率，都有一个被称为**日期偏移量**（date offset）的对象与之对应。例如，按小时计算的频率可以用 Hour 类表示。

有些频率所描述的时间点并不是均匀分隔的。例如，"M"（日历月末）和"BM"（每月最后一个工作日）就取决于每月的天数，对于后者，还要考虑月末是不是周末。由于没有更好的术语，我将这些称为**锚点偏移量**（anchored offset）。

2.3.3 移动（超前和滞后）数据

移动（shifting）指的是沿着时间轴将数据前移或后移。Series 和 DataFrame 都有一个 shift 方法用于执行单纯的前移或后移操作，保持索引不变。

2.4 时区处理

2.4.1 时区本地化和转换

默认情况下，pandas中的时间序列是单纯的（naive）时区。从单纯到本地化的转换是通过 `tz_localize` 方法处理。一旦时间序列被本地化到某个特定时区，就可以用 `tz_convert` 将其转换到别的时区。

2.4.2 操作时区意识型 Timestamp 对象

跟时间序列和日期范围差不多，独立的 Timestamp 对象也能被从单纯型（naive）本地化为时区意识型（time zone-aware），并从一个时区转换到另一个时区。

2.4.3 不同时区之间的运算

如果两个时间序列的时区不同，在将它们合并到一起时，最终结果就会是 UTC。由于时间戳其实是以 UTC 存储的，所以这是一个很简单的运算，并不需要发生任何转换。

Table 2.3: 基本时间序列频率（不完整）

别名	偏移类型	描述
D	Day	日历日报
B	BusinessDay	商业日报
H	Hour	每小时
T或者min	Minute	一分钟一次
S	Second	每秒一次
L或者ms	Milli	毫秒（1 秒的 1/1,000）
U	Micro	微秒（1 秒的 1/1,000,000）
M	MonthEnd	每月最后一个日历日
BM	BusinessMonthEnd	每月最后一个工作日（工作日）
MS	MonthBegin	每月的第一个日历日
BMS	BusinessMonthBegin	每月第一个工作日
W-MON, W-TUE, ...	Week	每周的指定日期（周一、周二、周三、周四、周五、周六或周日）
WOM-1MON, WOM-2MON, ...	WeekOfMonth	生成每月第一周、第二周、第三周或第四周的每周日期（例如，WOM-3FRI每个月的第三个星期五）
Q-JAN, Q-FEB, ...	QuarterEnd	以指定月份结束的年份的季度日期固定在每个月的最后一个日历日（一月、二月、三月、四月、五月、六月、七月、八月、九月、十月、十一月或十二月）
BQ-JAN, BQ-FEB, ...	BusinessQuarterEnd	对于以指定月份结束的年份，季度日期固定在每月的最后一个工作日
QS-JAN, QS-FEB, ...	QuarterBegin	对于以指定月份结束的年份，季度日期固定在每个月的第一个日历日
BQS-JAN, BQS-FEB, ...	BusinessQuarterBegin	对于以指定月份结束的年份，季度日期固定在每月的第一个工作日
A-JAN, A-FEB, ...	YearEnd	年度日期固定在给定月份的最后一个日历日（一月、二月、三月、四月、五月、六月、七月、八月、九月、十月、十一月或十二月）
BA-JAN, BA-FEB, ...	BusinessYearEnd	年度日期固定在给定月份的最后一个工作日
AS-JAN, AS-FEB, ...	YearBegin	年度日期固定在给定月份的第一天
BAS-JAN, BAS-FEB, ...	BusinessYearBegin	年度日期固定在给定月份的第

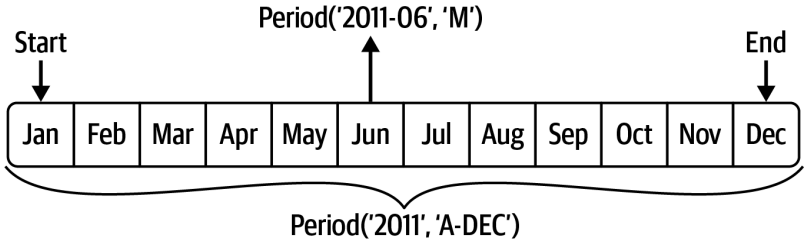

Figure 2.1: Period frequency conversion illustration

Figure 2.2: Different quarterly frequency conventions

2.5 时期及其算术运算

时期（period）表示的是时间区间，比如数日、数月、数季、数年等。Period 类所表示的就是这种数据类型，其构造函数需要用到一个字符串或整数，以及 **??** 中的频率。

只需对 Period 对象加上或减去一个整数即可达到根据其频率进行位移的效果。如果两个 Period 对象拥有相同的频率，则它们的差就是它们之间的单位数量。

2.5.1 时期的频率转换

2.5.2 按季度计算的时期频率

季度型数据在会计、金融等领域中很常见。许多季度型数据都会涉及“财年末”的概念，通常是一年 12 个月中某月的最后一个日历日或工作日。就这一点来说，时期”2012Q4”根据财年末的不同会有不同的含义。pandas 支持 12 种可能的季度型频率，即 Q-JAN 到 Q-DEC。

2.5.3 将Timestamp转换为Period（及其反向过程）

通过使用 `to_period` 方法，可以将由时间戳索引的 Series 和 DataFrame 对象转换为以时期索引。由于时期指的是非重叠时间区间，因此对于给定的频率，一个时间戳只能属于一个时期。新 PeriodIndex 的频率默认是从时间戳推断而来的，你也可以指定任何别的频率。结果中允许存在重复时期。要转换回时间戳，使用 `to_timestamp` 即可。

2.5.4 通过数组创建 PeriodIndex

2.6 重采样及频率转换

重采样（resampling）指的是将时间序列从一个频率转换到另一个频率的处理过程。将高频率数据聚合到低频率称为降采样（downsampling），而将低频率数据转换到高频率则称为升采样（upsampling）。并不是所有的重采样都能被划分到这两个大类中。例如，将 W-WED（每周三）转换为 W-FRI 既不是降采样也不是升采样。

pandas 对象都带有一个 resample 方法，它是各种频率转换工作的主力函数。resample 有一个类似于 groupby 的 API，调用 resample 可以分组数据，然后会调用一个聚合函数。

2.6.1 升采样和插值

在将数据从低频率转换到高频率时，就不需要聚合了。

2.6.2 通过时期进行重采样

升采样要稍微麻烦一些，因为你必须决定在新频率中各区间的哪端用于放置原来的值，就像 asfreq 方法那样。convention 参数默认为 'start'，也可设置为 'end'。

由于时期指的是时间区间，所以升采样和降采样的规则就比较严格：

- 在降采样中，目标频率必须是源频率的子时期（subperiod）。
- 在升采样中，目标频率必须是源频率的超时期（superperiod）。

Table 2.4: resample method arguments

参数	描述
rule	指示所需重采样频率的字符串、DateOffset 或 timedelta（例如，“M”、“5min”或Second(15)）
axis	重新采样的轴；默认axis=0
`fill_method`	上采样时如何插值，如”ffill”或”bfill”；默认情况下不进行插值
closed	在下采样中，每个间隔的哪一端是封闭的（包括），”right”或者”left”
label	在下采样中，如何使用”right”或”left”bin 边缘标记聚合结果（例如，可以将 9:30 到 9:35 的五分钟间隔标记为9:30或9:35）
limit	向前或向后填充时，填充的最大周期数
kind	聚合到周期 (“period”) 或时间戳 (“timestamp”)；默认为时间序列的索引类型
convention	重采样周期时，将低频周期转换为高频周期的约定（”start”或”end”）；默认为”start”
origin	用于确定重采样 bin 边缘的“基本”时间戳；“epoch”也可以是、“start”、“`start_day`”、”end”、或“`end_day`”；resample有关完整详细信息，请参阅文档字符串
offset	添加到原点的偏移时间增量；默认为None

Chapter 3

Appendix A