Timeseries Data in Pandas

30 Pandas 时间序列数据

重塑数组的维数、形状

很难做出预测，尤其是对未来的预测。

It is difficult to make predictions, especially about the future.

—— 尼尔斯·玻尔 (Niels Bohr) | 丹麦物理学家 | 1885 ~ 1962

30.1 什么是时间序列?

时间序列 (timeseries) 是指按照时间顺序排列的一系列数据点或观测值，通常是等时间间隔下的测量值，如每天、每小时、每分钟等。时间序列数据通常用于研究时间相关的现象和趋势，例如股票价格、气象数据、经济指标等。图 1 (a) 所示为标普 500 (S&P 500) 数据。

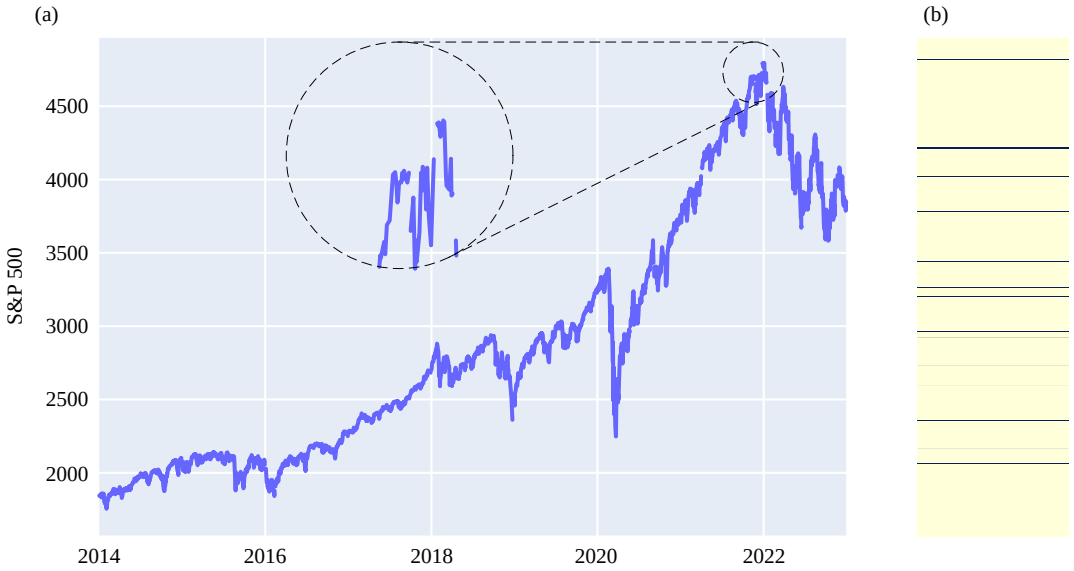

图 1. 标普 500 数据，含有缺失值

时间序列分析是一种重要的数据分析方法，它可以用于预测未来的趋势和变化，评估现有趋势的稳定性和可靠性，并发现异常点和异常趋势。时间序列分析通常包括以下几个步骤：

- 数据预处理：对数据进行清洗、去噪、填补缺失值等操作，以提高数据质量和可靠性。
- 时间序列的可视化：对数据进行绘图，以了解数据的分布、趋势和周期性。
- 时间序列的统计分析：对数据进行时间序列分解、平稳性检验、自相关性检验等统计分析，以评估数据的稳定性和相关性。
- 时间序列的建模和预测：根据统计分析的结果，建立合适的时间序列模型，进行未来趋势的预测和评估。

比如，在图 1 (a) 中被局部放大的曲线上，大家已经看到了缺失值。图 1 (b) 用热图可视化缺失值的位置。在本章配套的代码中，大家会看到经过计算缺失值的占比约为 3.5%。

本章仅仅采用“图解”介绍部分时间序列分析，《数据有道》一册将专门介绍时间序列相关话题。

Pandas 中的时间序列功能

在 Python 中，Pandas 库提供了强大的时间序列处理和分析功能，使得时间序列的处理和分析变得更加简单和高效。在 Pandas 中，时间序列分析的主要方法包括：

- 创建时间序列：可以通过 pandas.date_range() 方法创建一个时间范围，或者将字符串转换为时间序列对象。
- 时间序列索引：可以使用时间序列作为 DataFrame 的索引，从而方便地进行时间序列分析。
- 时间序列的切片和索引：可以使用时间序列的标签或位置进行切片和索引。
- 时间序列的重采样：可以将时间序列转换为不同的时间间隔，例如将日频率的数据转换为月频率的数据。
- 移动窗口函数：可以对时间序列数据进行滑动窗口操作，计算滑动窗口内的统计指标，例如均值、方差等。
- 时间序列的分组操作：可以将时间序列数据按照时间维度进行分组，从而进行聚合操作，例如计算每月的平均值、最大值等。
- 时间序列的聚合操作：可以对时间序列数据进行聚合操作，例如计算每周、每月、每季度的总和、平均值等。
- 时间序列的可视化：可以使用 Pandas、Matplotlib、Seaborn、Plotly 等库对时间序列数据进行可视化，例如绘制线形图、散点图、直方图等。

30.2 缺失值

缺失值 (missing value) 指的是数据集中的某些值缺失或未被记录的情况。它们可能是由于测量设备故障、记录错误、样本丢失或数据清洗不完整等原因导致的。缺失值可能在数据分析和建模中产生严重的影响，因为它们会导致数据样本的大小不一致，使得数据的统计分布和关系不准确或无法得出。另外，许多机器学习算法无法处理缺失值，必须对其进行处理或者删除。

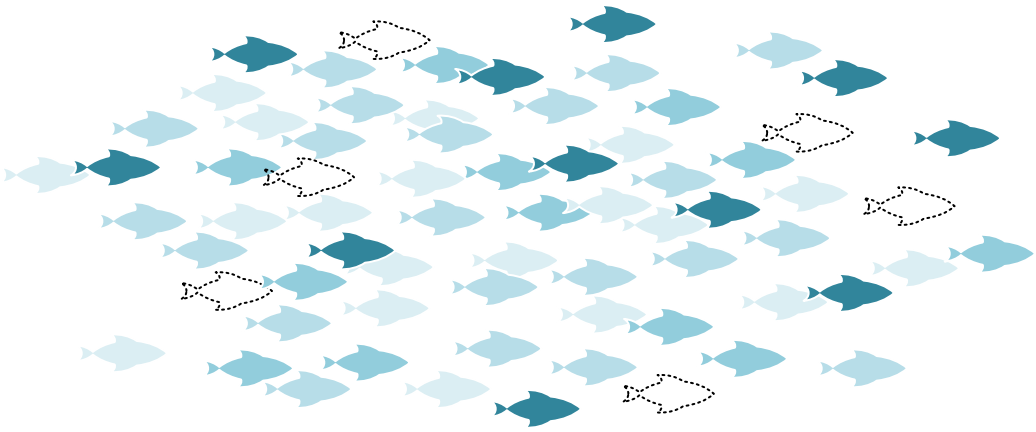

图 2. 缺失值

在数据处理中，通常需要对缺失值进行识别、处理或删除。一些处理缺失值的方法包括：

- 删除带有缺失值的样本或变量。
- 使用常量填充缺失值，例如用零、平均值、中位数等常量填充。
- 使用回归模型、插值方法等技术，对缺失值进行预测和填充。
- 对于分类变量，可以创建一个新的类别来表示缺失值。

在选择处理方法时，需要根据具体情况和数据分析的目的来决定。

图 1 中的缺失值则对应非营业日，比如周六日、节假日等。将这些缺失值删除之后，我们便得到图 3 所示的趋势。为了醒目地观察每年趋势，我们绘制了图 4。

鸢尾花书《数据有道》将介绍处理缺失值的各种方法。

图 3. 标普 500 数据，删除缺失值

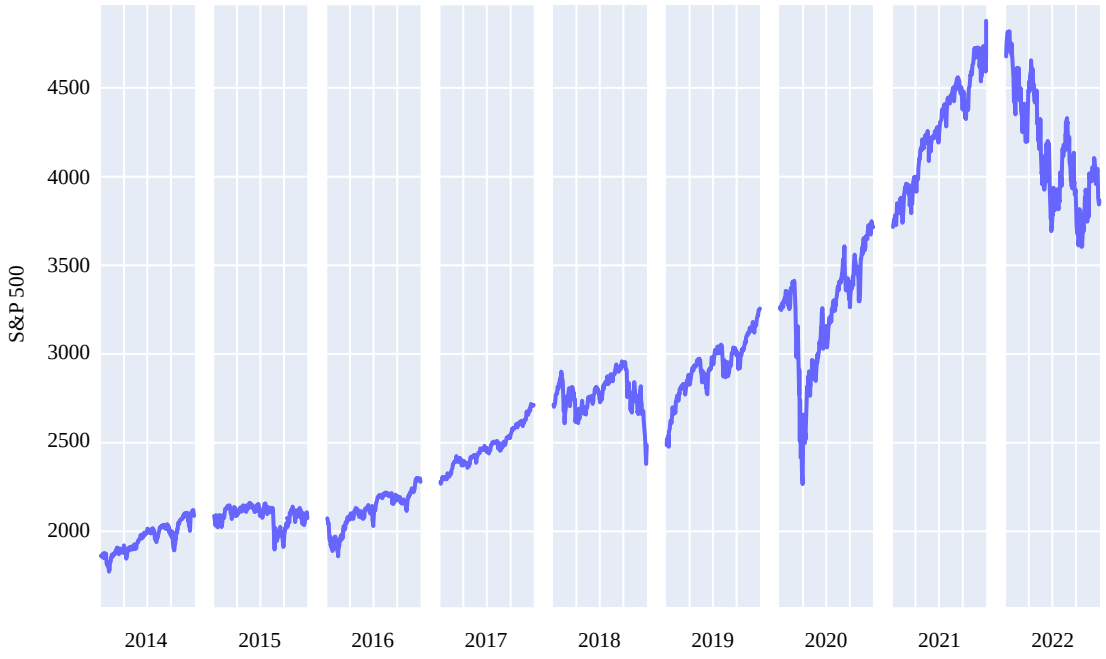

图 4. 标普 500 数据，按年观察趋势

什么是离群值？

在统计学和数据分析中，离群值 (outlier) 指的是在数据集中与其他数据值显著不同的异常值。它们可能是由于测量误差、实验异常、录入错误、样本损坏或数据处理错误等因素导致的。离群值具有比其他数据点更大或更小的数值，与其他数据点之间的差异通常非常显著。

离群值会对数据分析结果产生影响，比如对平均值、方差、相关性等统计指标的计算都会受到其影响。因此，在数据分析和建模中，需要对离群值进行识别、处理或删除。常见的方法包括使用箱线图或 3σ 准则等方法来识别离群值，并根据具体情况进行处理或删除。如果离群值确实是数据中真实存在的异常值，则可能需要对其进行单独分析或建立针对其的模型。

本章不会介绍如何处理离群值，相关内容请参考《数据有道》。

30.3 移动平均

时间序列的移动平均 (moving average, MA) 是一种常用的平滑技术，用于去除序列中的噪声和波动，以便更好地观察和分析序列的长期趋势。

移动平均通过计算序列中一段固定长度（通常称为窗口）内数据点的平均值来平滑序列。窗口的大小决定了平滑的程度，较大的窗口将平滑更多的波动，但可能会导致较长的滞后。

具体步骤如下：

- 1) 选择窗口的大小，例如 10 个数据点。
- 2) 从序列的起始位置开始，计算窗口内数据点的平均值。
- 3) 将该平均值作为移动平均的第一个数据点，记录下来。
- 4) 移动窗口向后滑动一个数据点的位置。
- 5) 重复步骤 2 至 4，计算新窗口内的平均值，并记录下来。
- 6) 继续滑动窗口直到到达序列的末尾，得到一系列移动平均值。

移动平均的计算可以使用简单移动平均 (Simple Moving Average, SMA) 或加权移动平均 (Weighted Moving Average, WMA) 来进行。简单移动平均对窗口内的每个数据点赋予相等的权重，而加权移动平均则可以根据需求赋予不同的权重，以更强调某些数据点的重要性。

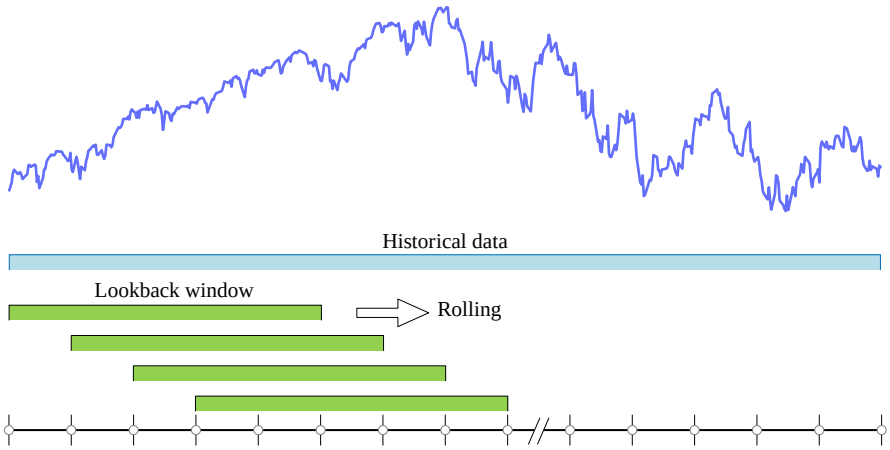

图 5. 移动窗口

通过计算移动平均，时间序列中的短期波动可以平滑，从而更容易观察到长期趋势和周期性变化。移动平均在金融分析、经济预测和数据分析等领域得到广泛应用。

图 6. 标普 500 数据，移动平均

30.4 收益率

为了量化股票市场的每日涨跌，我们需要计算股票的日收益率。计算当日收益率时需要知道两个关键数据点：股票的当日收盘价、前一日收盘价。

日收益率的计算公式为：日收益率 = (当日收盘价 − 前一日收盘价) / 前一日收盘价。将这个公式应用于具体的股票数据，就可以计算出每个交易日的日收益率。图 7 所示为标普 500 的日收益率。

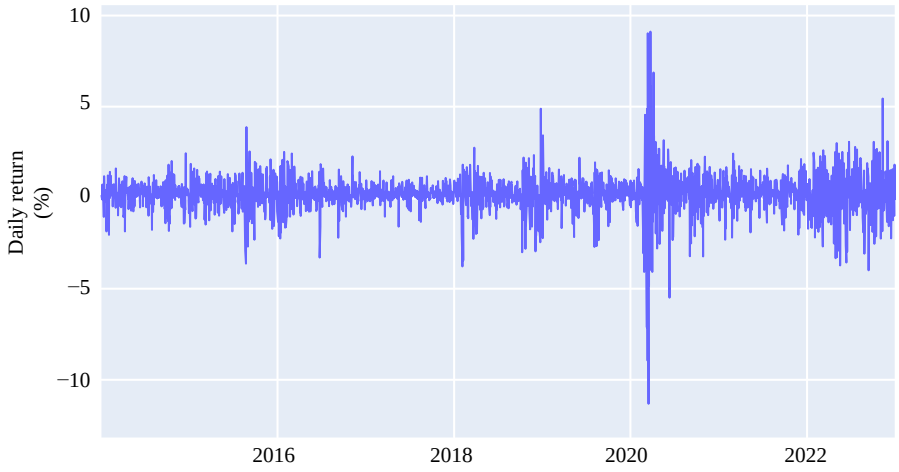

图 7. 标普 500 数据日收益率

为了更方便观察每年涨跌情况，我们绘制了图 8。

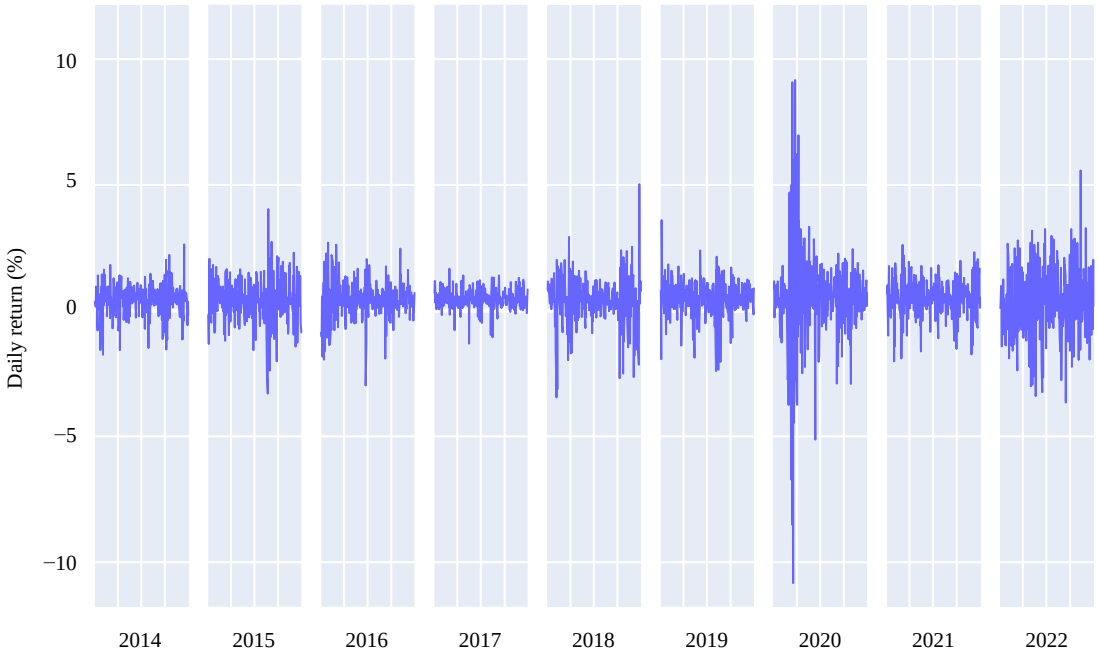

图 8. 标普 500 数据日收益率，按年观察趋势

30.5 统计分析

市场涨跌越越剧烈，曲线波动越剧烈。图 8 这些曲线类似随机行走，为了发现规律，我们需要借助统计工具。

年度分布

图 9 所示为下载所有数据计算得到日收益率绘制的分布图。大家可以从分布中计算得到均值和标准差。这个任务交给大家自行完成。图 10 所示为年度日收益率分布变化情况。

为了更好的量化股票的波动情况，我们需要一个指标——波动率 (volatility)。波动率是衡量其价格变动幅度的指标，常用的量化方法为历史波动率 (historical volatility)。历史波动率本质上就是一定回望窗口内收益率样本数据的标准差。

图 11 所示为利用水平柱状图可视化日收益率的年度均值、波动率 (标准差)。

此外，我们还可以使用山脊图 (ridgeline plot) 可视化每年收益率的分布情况，具体如图 12 所示。Joypy 是一个 Python 库，用于创建山脊图。山脊图是一种可视化工具，用于展示多个连续变量在一个维度上的分布，并且能够显示不同组之间的比较。

山脊图的特点是将多个曲线图，通常是核密度估计曲线，沿着一个共享的垂直轴线堆叠显示，形成一座山脉状的图形。每个曲线代表一个组或类别，可以通过颜色或其他视觉属性进行区分。要使用 Joypy 绘制山脊图，需要首先安装 Joypy 库，并导入 joyplot 模块。

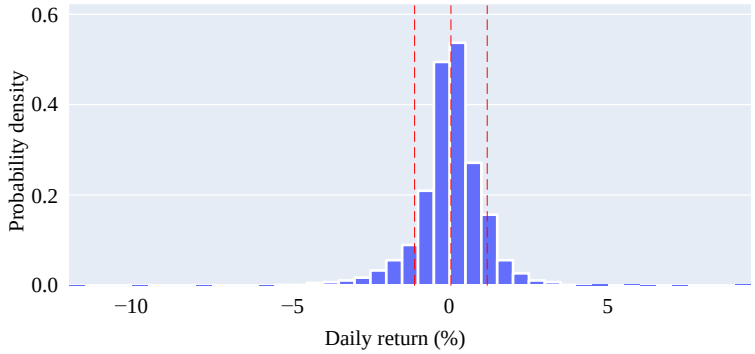

图 9. 所有下载历史数据日收益率分布

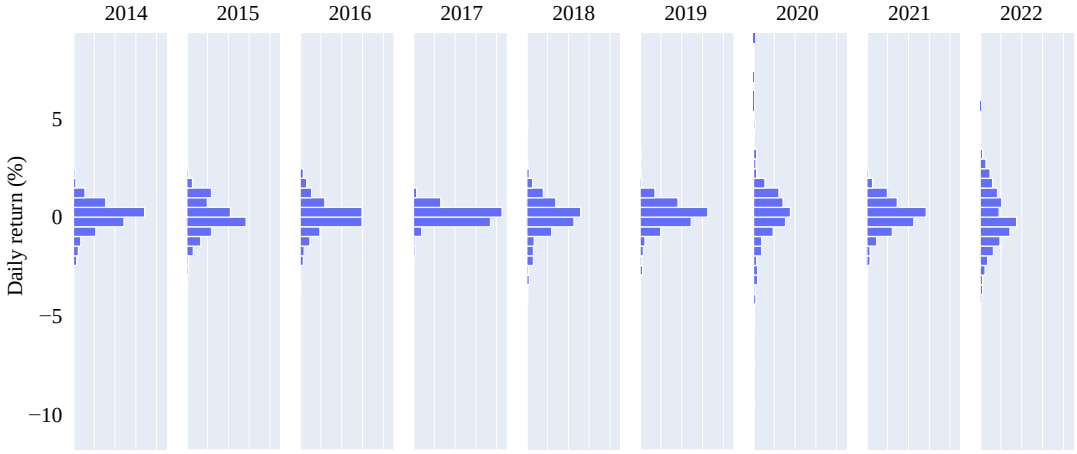

图 10. 日收益率分布，按年

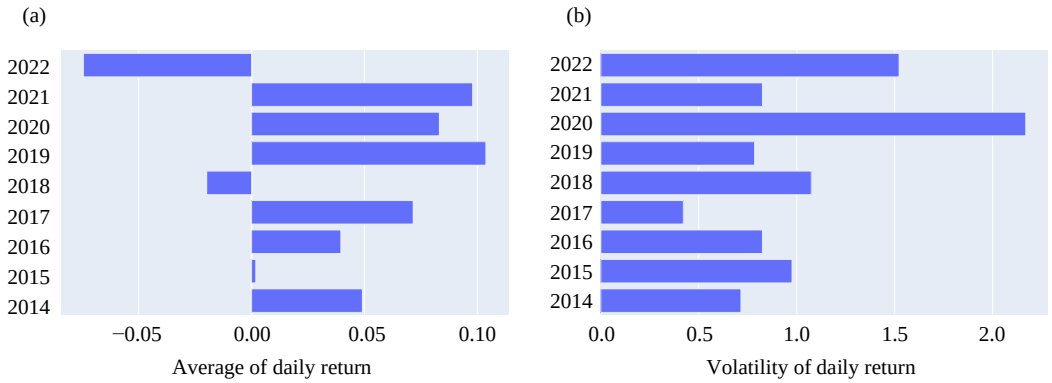

图 11. 水平柱状图可视化收益率均值、标准差 (波动率)，按年

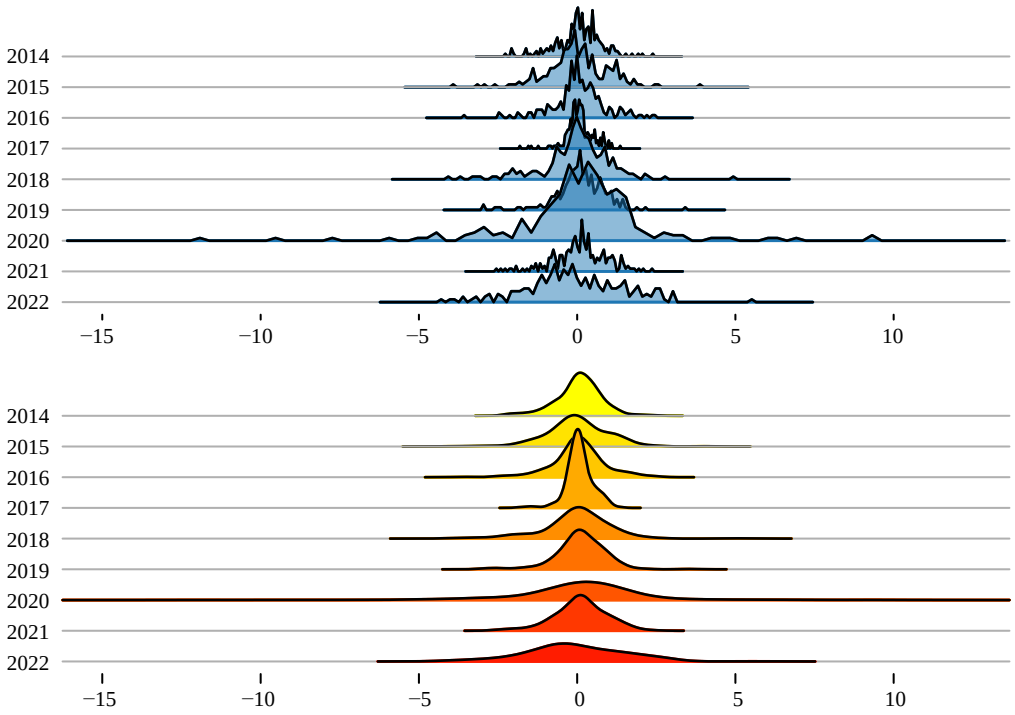

图 12. 山脊图，按年

季度分布

当然，我们也可以按季度分析收益率。图 13 所示为每个季度收益率的均值、标准差的柱状图。图 14 所示为每个季度收益率的山脊图。这幅图我们把纵轴的时间隐去。

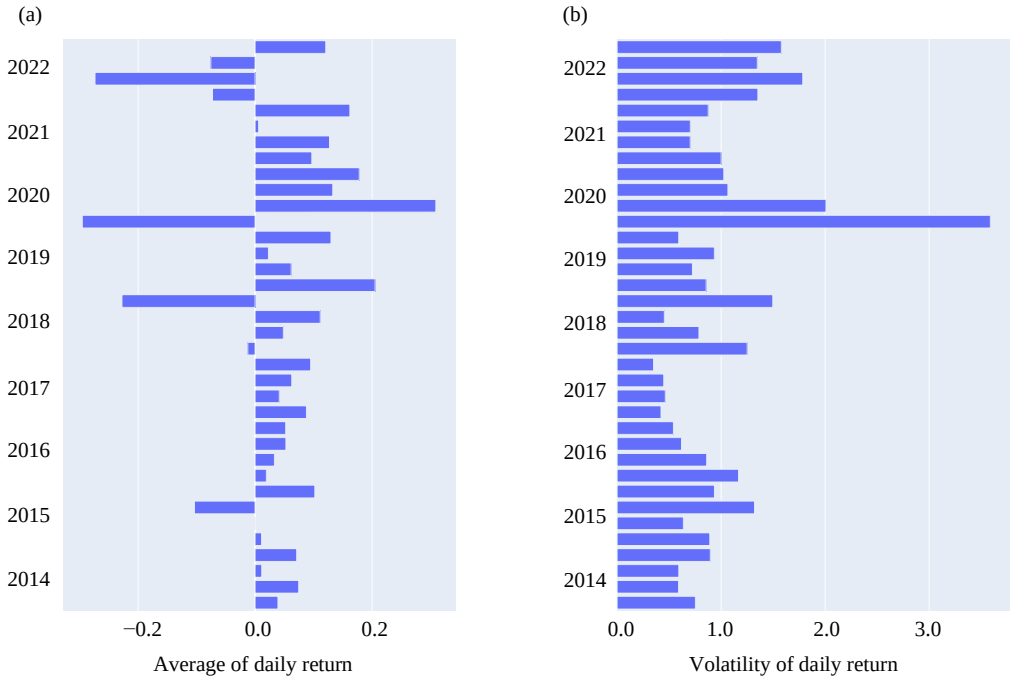

图 13. 水平柱状图可视化收益率均值、标准差 (波动率)，按季度

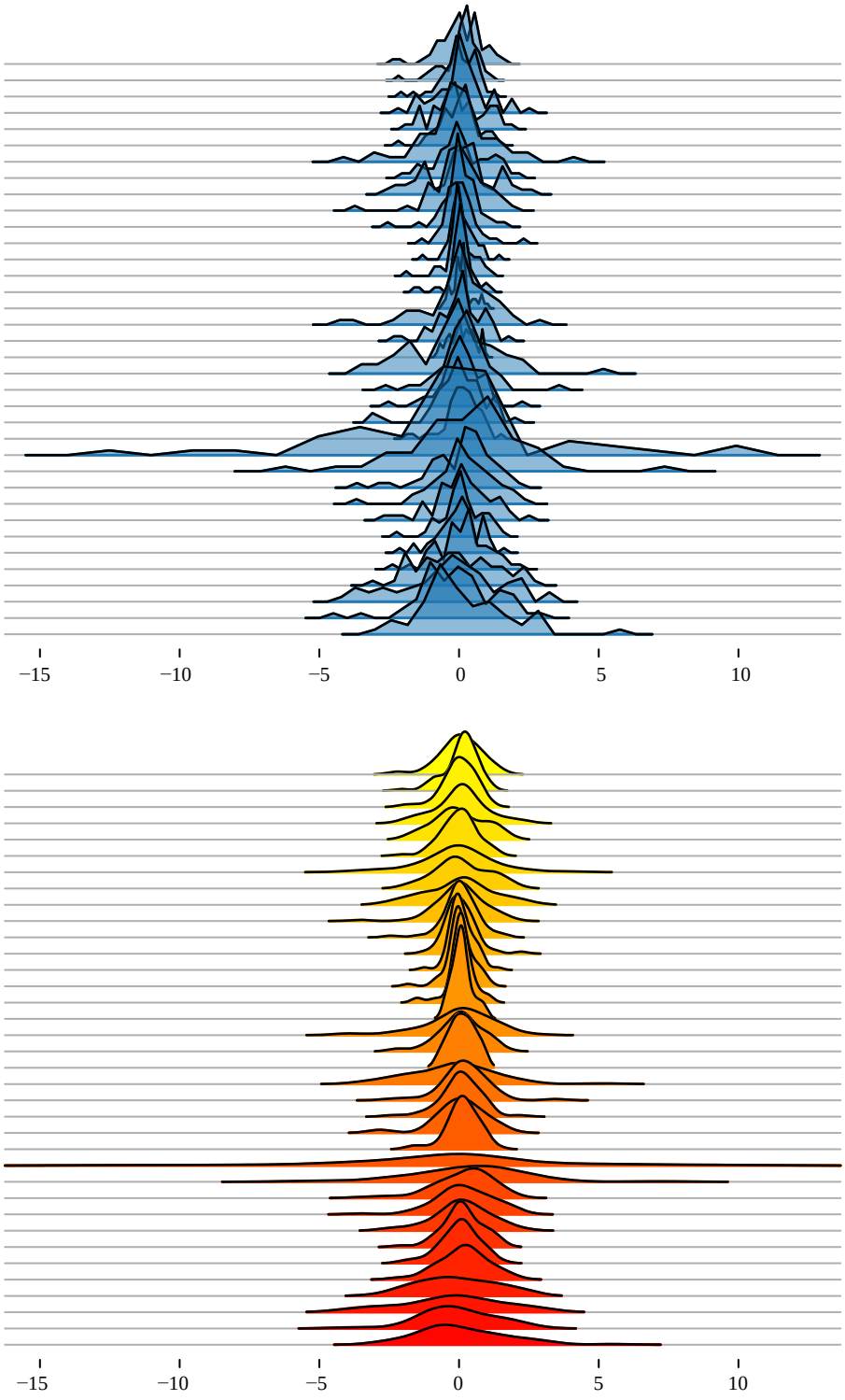

图 14. 山脊图，按季度

移动波动率

准确来说，历史波动率是根据过去一段时间内的股票价格数据计算得出的波动率。

可以选择一个时间窗口，例如 20 天 (一个月)、60 天 (一个季度)、125 或 126 天 (半年)、250 或 252 天 (一年)，计算每个交易日的收益率，然后求得其标准差，最终得到历史波动率。当这个回望窗口移动时，我们便得到移动波动率的时间序列数据。图 15 所示的移动波动率的回望窗口长度为 250 天 (营业日)。请大家自己修改回望窗口长度 (营业日数量)，比较移动波动率曲线。

《数据有道》还会专门介绍指数加权移动平均 EWMA 方法计算的均值和波动率。

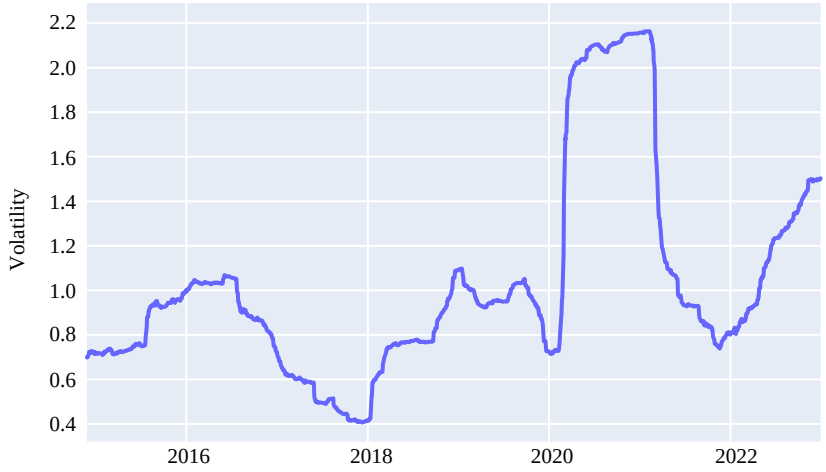

图 15. 移动波动率

30.6 相关性

几个不同时间序列之间肯定也会存在相关性。图 16 所示为标普 500 日收益率和三个汇率收益率之间的相关性系数矩阵热图。

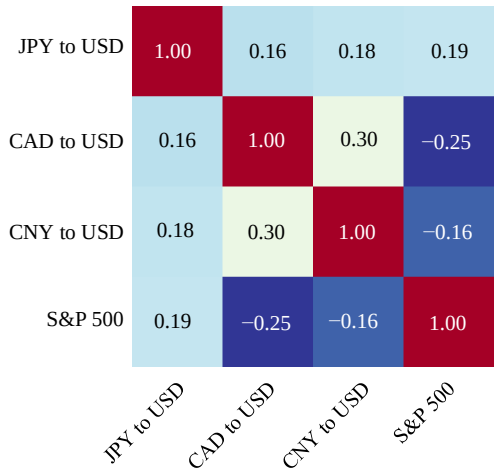

图 16. 相关性系数矩阵

相关性并不是一成不变的，也是随时间不断变化。如图 17 所示，当我们指定具体的移动窗口长度，也可以计算移动相关性。

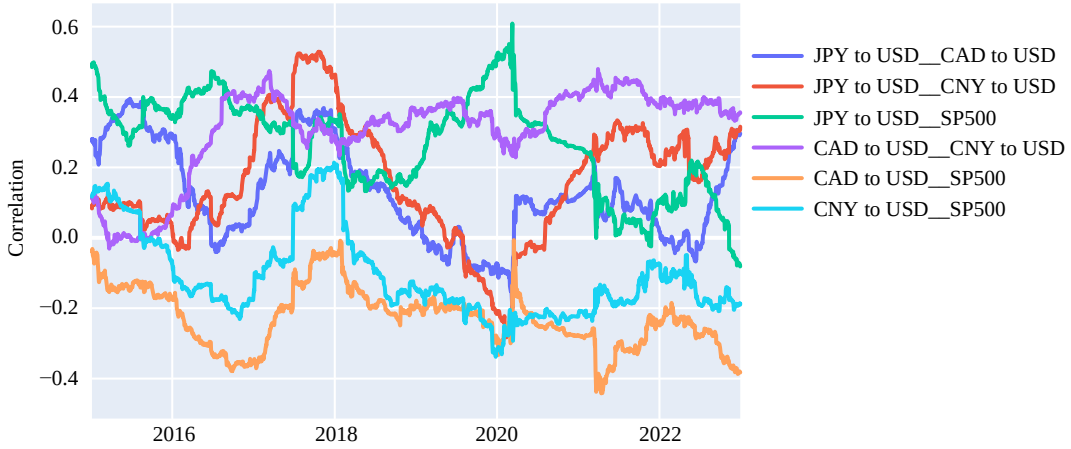

图 17. 移动相关性

对时间序列历史数据完成分析后自然少不了预测这个环节。本书不会展开讲解，请大家参考《数据有道》。

请大家完成下面这道题目。

Q1. 请大家把本章配套代码中历史数据截止时间修改为最近日期，重新下载数据逐步完成本章前文时间序列分析。

* 本章不提供答案。

有关 Pandas 中时间序列更多用法，请大家参考：

https://pandas.pydata.org/docs/user_guide/timeseries.html

此外，Statsmodels 有大量时间序列分析工具：

https://www.statsmodels.org/stable/user-guide.html#time-series-analysis