00Pandas介绍

Pandas是 Python 语言开发的用于数据处理（data manipulation）和数据分析（data analysis）的第三方库。它擅长处理数字型数据和时间序列数据，当然也文本型的数据也能轻松处理。

官方介绍如下：

> *Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.*

一、名字来源

Pandas 的命名来源并非「熊猫」，而是来自于计量经济学中术语**面板数据**（Panel data），它是一种数据集的结构类型，具有横截面和时间序列两个维度。不过，我们不用必须了解它，它只是一种灵感、思想来源。

二、用途

那么问题来了：numpy已经能够帮助我们处理数据，能够结合matplotlib解决我们数据分析的问题，那么pandas学习的目的在什么地方呢？

numpy能够帮我们处理处理数值型数据，但是这还不够， 很多时候，我们的数据除了数值之外，还有字符串，还有时间序列等。比如：我们通过爬虫获取到了存储在数据库中的数据。

Pandas 对数据的处理是为数据的分析服务的，它所提供的各种数据处理方法、工具是基于数理统计学出发，包含了日常应用中的众多数据分析方法。我们学习它不光掌控它的相应操作技术，还要从它的处理思路中学习数据分析的理论和方法。

特别地，想成为或者转行数据分析师、数据产品经理、数据开发等和数据相关工作者的同学，学习 Pandas 更能让你深入数据理论和实践，更好地理解和应用数据。

Pandas 可以轻松应对白领们日常工作中的各种表格数据处理，还应用在金融、统计、数理研究、物理计算、社会科学、工程等领域里。

Pandas 可以实现复杂的处理逻辑，这些往往是 Excel 等工具无法处理的，还可以自动化、批量化，对于相同的大量的数据处理我们不需要重复去工作。

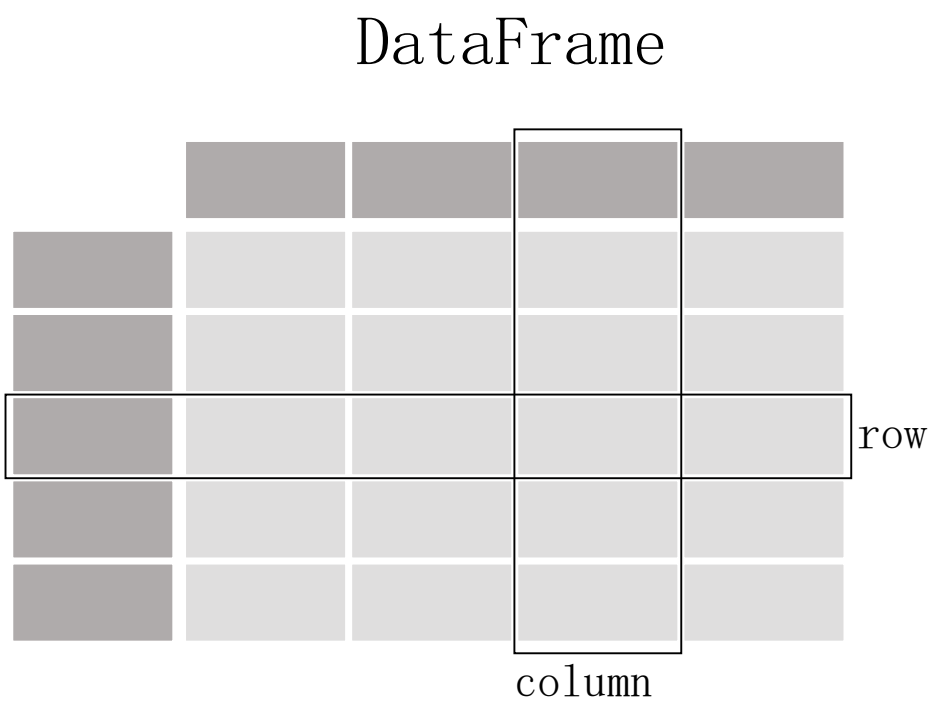
Pandas 可以做到非常震撼的可视化，它对接众多的高颜值可视化库，可以实现动态数据交互效果。

以上这些，在往后的学习和工作中，你会有所体会。

三、适用的数据

Pandas 适合处理一个规正的二维数据（一维也可以，应用较少），即有 `N 行 N 列`，类似于 `SQL` 执行后产出的，或者 `无合并单元格Excel 表格` 这样的数据。它可以把多个文件的数据合并在一起，如果结构不一样，也可以经过处理进行合并。

这里说的二维数据是指，像一个矩形的平面在横向和纵向被分隔成多个格子，每个格子里存放一个数据。



上图是一个 pandas 中定义的数据框架。

四、基本功能

常用的基本功能有：

- 从 Excel、CSV、网页、SQL、剪贴板等读取数据
- 合并多个文件或者 sheet 数据，拆分数据为独立文件
- 数据清洗，如去重、缺失值、填充默认值、格式补全、极端值处理等
- 建立高效的索引
- 支持大体量数据
- 按一定业务逻辑插入计算后的列、删除列
- 灵活方便的数据查询、筛选
- 分组聚合数据，可独立指定分组后的各字段计算方式
- 数据的转置，如行转列列转行变更处理
- 连接数据库，直接 SQL 查询数据并进行处理
- 对时序数据进行分组采样，如按月、按季、按工作小时，也可以自定义周期，如工作日
- 窗口计划，移动窗口统计、日期移动等
- 灵活的可视化图表输出，支持所有的统计图形
- 融合在表格的样式风格，提高数据识别效率

等等。

五、学习方法

对于一个新的工具，从我们的目标出发就是能够使用它，让它发挥价值。因此，最好的方法就拿一个自己熟悉的数据去处理它，同时把日常工作需要手工处理的表格用 Pandas 来做，刚开始可能不能完全替代，但随时慢慢积累，就会得心应手。

在学习初期，只需要对着教程去模仿，把涉及到的常用操作总结归纳。养成遇到不懂的查看函数说明和查官方文档「https://pandas.pydata.org/docs/」的习惯。

本课程侧重点在 Pandas 的使用上面，暂不过多地讲解数据分析方法，不过 Pandas 提供的数据分析方法就是给我们提供了一个数据分析思路，可以帮助我们建立完善数据分析理论体系。