Introduction to Pandas

WISER CLUB
钱晨

Content

- What is pandas
- How to analyze data with Pandas
- Some examples

What is Pandas

- Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.

- Python中用于处理数据以及进行数据分析的模块具有高效、简洁等特点

推荐学习资料

- 利用Python进行数据分析—-Wes McKinney著
    - 入门，建议阅读5到10章
- 利用Python进行科学计算
- Official Documentation
    - http://pandas.pydata.org/pandas-docs/stable/index.html

Basic IO tools

- 读写csv, xlsx等文件

- 连接数据库，如Oracle,SQL Server等

Basic IO tools

- 读写CSV文件
- pandas.read_csv(filepath, header, names, skiprows, encoding)
  - header: int(指CSV文件中的第几行)
  - names： arrary like(list等）如果header=None,接受names传入
  - skiprows: int
  - encoding:常见的为utf-8, gbk, gb18030, gb2312, iso-8859-1

Basic IO tools

- 读写xlsx文件

- pandas.read_excel(filepath, sheeetname, header, names, skiprows)

Intro to Data Structures

- Series:
    - It is a one-dimensional array
    - It is available to holding different data types and data labels(index)
- DataFrame
    - It is a kind of array with two dimensions
    - It is very similar to the data frame in R programming
- Panel(not included in)
    - Still developing

Series

- 创建Series：接受dict, list, array等对象
- Series的基本处理（如算数运算等）
- Series的属性

DataFrame

- 创建DataFrame:
    - 直接创建或者通过IO读写
- Column Selection，addition and deletion
- Basic attributes of data frame

Visualization

- pandas shows a great combination with the professional drawing tools——matplotlib which we we will discuss in the following lectures.

- In this part, we will show how to use pandas module to draw the ggplot style plots.

Visualization

- Plot
- Boxplot and violin plot
- scatter plot and hexbin plot

Merge, join and concat

- Concat：
    - deals with the heavy lifts of operation among different pandas objects along specific axis
- Merge, join
    - Database-style data frame joining methods

Working with Missing Data

- Cleaning and filling missing data
    - Replace the NA value with a scalar value
    - Fill gaps forward or backward
    - Limit the amount of filling
    - Filling with a pandas object

Computational tools

- Basic statistics function and methods
- Correlation, covariance, skew, mean and etc.

Computational tools

- Statistics functions
  - pandas对象都支持NumPy的数组接口，可以直接使用numpy的ufunc函数进行计算
  - 此外pandas提供基本的运算方法如mean, std, correlation, covariance等
    - 一般包括如下三个参数：axis, level, skin

Working with Missing Data

- Dropping axis labels with missing data
- Some interpolation methods
- Replacing generic values: scalar value, sequence, string or the regular expression.

Merge, join and concat

- Pandas provides an efficient way to combine different pandas object—-series and data frame.
- These kinds of methods will show a great edge over the manipulations in R programming.