自学报告：Pandas库

李帅 2016013270

周展平 2016013253

本文档说明：

本文档是针对Pandas 0.23.4的学习报告，主要是从实际的数据分析场景出发，以各个环节为线索学习Pandas的特性。

一、Pandas 简介

Pandas 是一个 Python 的开源项目，是数据分析的一个常用的工具。官方的文档中如下描述它：

> `pandas` is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

(linkage：http://pandas.pydata.org/pandas-docs/stable/overview.html)

Pandas 数据结构的实现基于Numpy，可视化方面则基于 Matplotlib，因此 Pandas 对于它们均有很好的兼容性。

二、基本数据类型

三、文件读写

四、数据索引(index)、排序(sort)

五、数据分组(group)

数据分组的含义包含3个方面：

1. **Split**：将数据按照一定的标准 (criterion) 进行分类。
2. **Apply**：对于每一个类别的数据，进行特定的操作。其中包括：
   （1）**Aggregation**：计算类别内数据的总体特征，如和、均值、维数
   （2）**Transformation**：对类内数据总体进行处理，如标准化、填补NA
   （3）**Filtration**：对某些类别的数据进行丢弃、筛选等
3. **Combine**：将经过操作的所有类别的数据重新按照某种方式组合起来 。

下面简单介绍具体的方法：

1. Split

六、合并(merge,join,concatenate)

七、可视化(visualize)

1、散点图 (plot)： 直接调用 matplotlob 的plot() 方法

八、参考资料