Chapter I Preparation

1 主要内容

虽然标题是“数据分析”，但是重点在于Python编程、库，以及用于数据分析的工具。

1.1 什么样的数据

主要指结构化数据（structured data），代指所有通用格式的数据，如：

- 表格型数据
- 多维数据（矩阵）
- 通过关键列（对于SQL，即主键及外键）相互联系的多个表
- 间隔平均or不平均的时间序列

2 重要的Python库

2.1 NumPy

NumPy - Numerical Python，科学计算的基础包，提供了以下功能（包括但不限于）：

- 快速高效的多维数组对象ndarray
- 用于对数组执行元素级计算以及直接对数组执行数学运算的函数

- 用于读写硬盘上基于数组的数据集的工具
- 线性代数运算、傅里叶变换，随机数生成

作为在算法和库之间传递数据的容器。

2.2 pandas

pandas提供了快速便捷处理结构化数据的大量数据结构和函数。

间距NumPy高性能数组计算功能以及电子表格和关系型数据库灵活的数据处理功能，本书重点。

2.3 matplotlib

最流行的用于绘制图表和其他二位数据可视化的Python库，默认的可视化工具

2.4 IPython和Jupyter

IPython变成了Jupyter庞大开源项目（一个交互和探索式计算的高效环境）中的一个组件。它最老也是最简单的模式，现在是一个用于编写、测试、调试Python代码的强化shell。你还可以使用通过Jupyter Notebook，一个支持多种语言的交互式网络代码“笔记本”，来使用IPython。IPython shell 和Jupyter notebooks特别适合进行数据探索和可视化。

Jupyter notebooks还可以编写Markdown和HTML内容，它提供了一种创建代码和文本的富文本方法。其它编程语言也在Jupyter中植入了内核，好让在Jupyter中可以使用Python以外的语言。

2.5 SciPy

专门解决科学计算中各种标准问题域的包的集合，主要包括下列包：

- scipy.integrate：数值积分例程和微分方程求解器。
- scipy.linalg：扩展了由numpy.linalg提供的线性代数例程和矩阵分解功能。
- scipy.optimize：函数优化器（最小化器）以及根查找算法。
- scipy.signal：信号处理工具。
- scipy.sparse：稀疏矩阵和稀疏线性系统求解器。
- scipy.special：SPECFUN（这是一个实现了许多常用数学函数（如伽玛函数）的Fortran库）的包装器。

- scipy.stats：标准连续和离散概率分布（如密度函数、采样器、连续分布函数等）、各种统计检验方法，以及更好的描述统计法。

…

3 安装及使用

安装anaconda就完事儿了。