# 《数据挖掘技术与应用》

|  |
| --- |
| 【实验名称】 实验：Pandas |
| 【实验目的】 |
| 1.了解Pandas模块的数据结构  2.使用Pandas对数据进行分析 |
| 【实验原理】 |
| Python Data Analysis Library 或 pandas 是基于NumPy 的一种工具，该工具是为了解决数据分析任务而创建的。Pandas 纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据集所需的工具。 |
| 【实验要求】 |
| 提交符合要求的实验报告：1.粘贴你补全的代码并注释原因。 2.将你程序运行的输出结果（图片）粘贴到报告中。3.回答STEP6中：什么是相关系数？和在STEP7中你发现的规律。 |
| 【实验步骤】 |
| **题目：使用Pandas对销售数据进行分析**现在我们获得了一个连续数据的数据集data.txt，该数据集由年，月份，和AB两地某商品的销售数据组成，我们希望发现这两地销售数据之间是否有关联。STEP1：查看数据集data.txt的数据构成，我们发现其数据有四列构成，分别代表的是年、月、以及两组销售数据。数据是和时间相关联的。因此我们以时间为索引，转换其数据结构将方便我们的分析。STEP2：构建convert\_data\_to\_timeseries函数，该函数目的是将原始数据中的某地销售数据转换为Series结构，其索引为年月构成的时间，因为Series强大的功能方便我们进行统计。问题一：完善代码实现convert\_data\_to\_timeseries函数功能，在这之前请先熟悉data\_range函数的使用方式，data\_range函数会返回一个时间索引。练习以下代码，查看其结果。 >>>date\_time\_index = pd.date\_range(start="2018-06-29", end="2018-07-02")  >>>date\_time\_index = pd.date\_range(start="2018-06-29", periods=4)  >>>date\_time\_index = pd.date\_range(periods=4, end="2018-07-02")  >>>date\_time\_index = pd.date\_range(start="2018-01-01", end="2018-07-01", freq='M')  >>>date\_time\_index = pd.date\_range(start="2018-01-01", end="2018-07-01", freq='D')  STEP3：查看数据，将A地（第三列）数据成图，查看其是否存在某种趋势    STEP4：加入第二组销售数据，将两组Series数据转化为二维数据结构DataFrame,索引为时间，DataFrame结构会利于我们的比较。问题二：补全代码，实现上述功能。    STEP5：查看数据，将AB地数据成图，查看两组数据否存在相同趋势      STEP6：从图像中似乎没发现数据的规矩，我们希望借助统计学来观测，问题三：分别计算两组数据的最大最小值，平均值，以及相关系数。请了解相关系数的函数corr()，以及什么是相关系数，请参考http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.corr.html  STEP7：将其相关系数的变化成图。请思考其是否存在规律？ |
|  |