实验报告

东北林业大学

信息与计算机科学技术实验中心

|  |
| --- |
| 一、实验目的  （1）掌握Pandas库中主要数据结构Series和DataFrame的主要操作，包括两类数据对象的创建、切片、添加行、列、修改元素、排序等方法；  （2）掌握利用Pandas库进行CSV和EXCEL文件的存取方法；  （3）掌握利用Pandas库对数据进行分组统计与透视图（交叉表）分析的方法。 |
| 二、实验环境  （1）硬件：PC机；  （2）软件：Anaconda Jupyter Notebook，Spyder，Pandas |
| 三、实验内容及结果  （一）Series和DataFrame数据对象的创建与基本操作  1、创建一个存放班级同学考试成绩的Series对象ser，索引为同学姓名，值为同学的《Python应用》课程的考试成绩，然后查询张晓梅同学的考试成绩。数据自拟即可，但要包含张晓梅的成绩。  2、查看题1中创建的序列对象的元素个数，索引值，形状。  3、对题1中创建的序列对象按照考试成绩降序排列后输出。  4、在题1中所创建的序列对象中添加李丽同学的考试成绩90后输出序列对象中后三个同学的成绩。  5、利用下表中数据创建一个DataFrame对象。   |  |  |  |  | | --- | --- | --- | --- | | 姓名 | 平时成绩 | 阶段成绩 | 期末成绩 | | 李明 | 100 | 80 | 100 | | 张凯 | 90 | 100 | 90 | | 刘梅 | 100 | 100 | 80 |   6、输出数据对象的维度、元素个数、形状和各列的数据类型。  7、输出所有同学的期末成绩和阶段成绩。  8、输出期末成绩在90分及以上的同学的所有成绩信息。  9、输出期末成绩在90分及以上的同学的阶段成绩和期末成绩。  10、添加王伟光同学的成绩，平时成绩90，阶段成绩70，期末试成绩80。然后删除张凯的成绩后，输出数据对象的值。  11、计算各位同学的总成绩，并进行输出，各部分成绩比例为平时20%，阶段30%，期末50%。  12、将姓名列设置为索引列，然后查询刘梅和王伟光的考试成绩。  （二）利用pandas相关函数对土地利用数据集进行数据分析  1、导入landuse.csv文件中的数据，并输出数据集中的前五行和后五行。  提示：利用pandas的read\_csv函数可以读取csv文件数据  2、查看数据的形状、元素个数、各列的数据类型。  3、查看数值型数据的统计特征，提示可以使用Dataframe的decribe( )方法。  4、查看主要有哪些地区，哪些县市；  5、查看各地区1980年的耕地面积，要求结果以耕地面积降序排列，以平方公里为单位，保留小数点后两位数字。提示：可以先选择出1980年地类编码>100的记录，然后根据地区名称进行分组后，求取各地区面积的总和，然后按照面积从大到小的顺序进行排序后输出。  6、计算各图斑1980年和2010年的一级土地利用代码。  7、利用透视表分析各地区2010年耕地、林地、草地、水域、建设用地、未利用地的面积。  （1）透视表实现方式：  （2）交叉表实现方式：  8、将统计的结果保存到excel文件中,文件名为“各县市土地利用面积统计表”。  9、将各地区内部不同显示的土地利用数据分别统计，保存到以不同地区命名的文件中，如哈尔滨市保存到stat文件夹下“哈尔滨市各县市土地利用面积统计表.xls”。  执行结果为：  10、分析哈尔滨市30年来的一级地类的土地利用变化情况，用土地利用变化矩阵显示出来。  11、上题目结果中的最后一列为1980年各类用地的面积，最后一行是2010年各类用地的面积，可以将这两部分数据进行合并，得到哈尔滨市两个时相各类用地面积对比数据。 |

|  |
| --- |
| 四、实验过程分析与讨论  1． 实验过程分析与讨论：  （1）pd.Series是pandas库中的一个函数，用于创建Series对象。在这里，index参数指定了Series对象的索引，data参数指定了Series对象的值。ser.size：这个属性返回序列中的元素数量。它告诉我们序列的长度或大小。ser.index：这个属性返回序列的索引。索引是用来标识序列中每个元素的标签或位置。它可以是整数、字符串或其他数据类型。ser.shape：这个属性返回序列的形状。对于一维序列，形状是一个整数，表示序列的长度。对于多维序列，形状是一个元组，包含每个维度的长度。  （2）sort\_values是排序函数，而ascending=False则是指定排序方式为降序排序。也就是说，这段代码会将ser中的数据按照从大到小的顺序进行排序，并返回排序后的结果。代码使用df.drop(1)删除df中索引为1的行，并将结果重新赋值给df。  （3）在pandas中head()函数可以查看前五行，tail()函数可以查看后五行，df.dtypes 是一个 Pandas DataFrame 对象的属性，用于返回该 DataFrame 中每列的数据类型。具体来说，它返回一个 Series 对象，其中包含了 DataFrame 中每列的名称作为索引，而对应的值则是该列的数据类型。  （4）df.describe() 是一个用于描述数据框（DataFrame）的方法。它提供了关于数据的统计摘要信息，包括计数、均值、标准差、最小值、25%分位数、中位数、75%分位数和最大值。具体来说，df.describe() 返回一个新的数据框，其中包含了每个数值列的统计摘要。  （5）由于pandas函数在更新，有些函数已经无法使用，不能在使用Series |
| 五、指导教师意见    指导教师签字：  年 月 日 |