**实验五 基于Pandas的数据分析**

**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、实验仪器及实验环境**

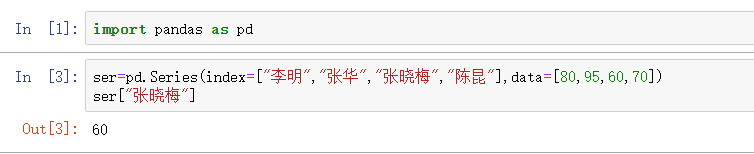
（1）硬件：PC机；

（2）软件：Anaconda Jupyter Notebook，Spyder，Pandas

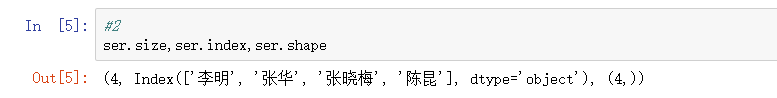
**三、实验内容：**

（一）Series和DataFrame数据对象的创建与基本操作

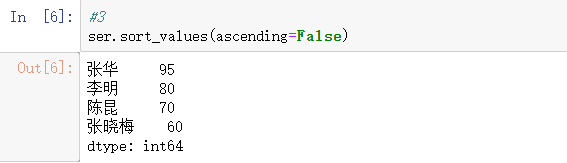
1、创建一个存放班级同学考试成绩的Series对象ser，索引为同学姓名，值为同学的《Python应用》课程的考试成绩，然后查询张晓梅同学的考试成绩。数据自拟即可，但要包含张晓梅的成绩。



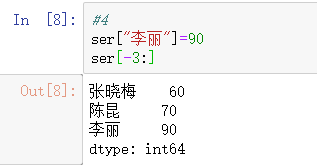
2、查看题1中创建的序列对象的元素个数，索引值，形状。



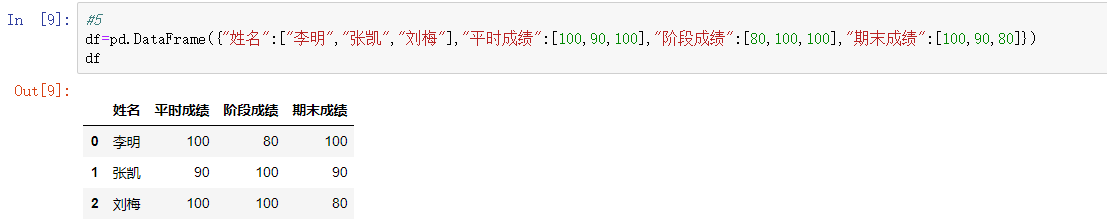
3、对题1中创建的序列对象按照考试成绩降序排列后输出。



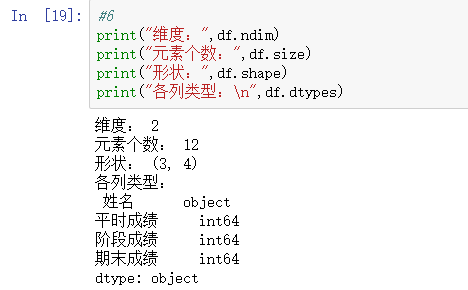
4、在题1中所创建的序列对象中添加李丽同学的考试成绩90后输出序列对象中后三个同学的成绩。

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。



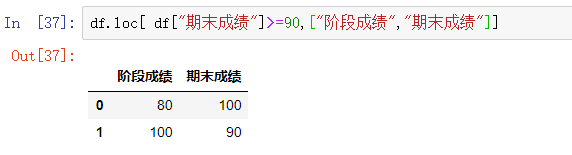
7、输出所有同学的期末成绩和阶段成绩。



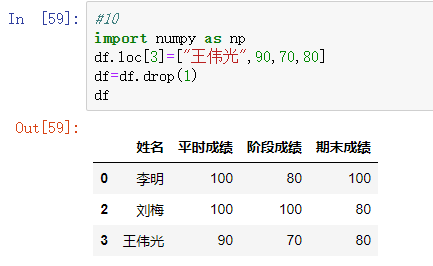
8、输出期末成绩在90分及以上的同学的所有成绩信息。



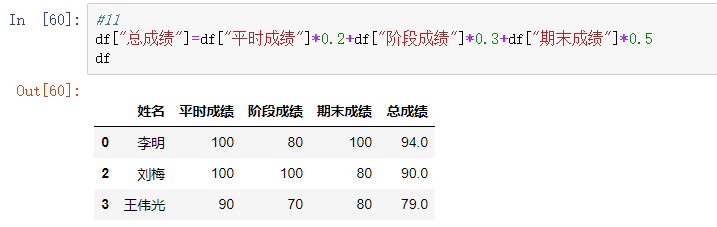
9、输出期末成绩在90分及以上的同学的阶段成绩和期末成绩。



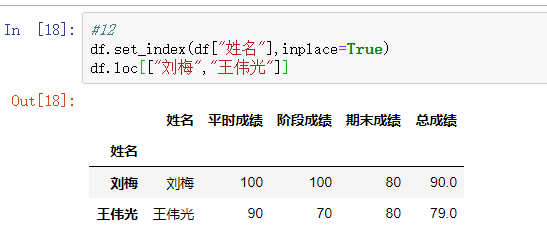
10、添加王伟光同学的成绩，平时成绩90，阶段成绩70，期末试成绩80。然后删除张凯的成绩后，输出数据对象的值。



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%。



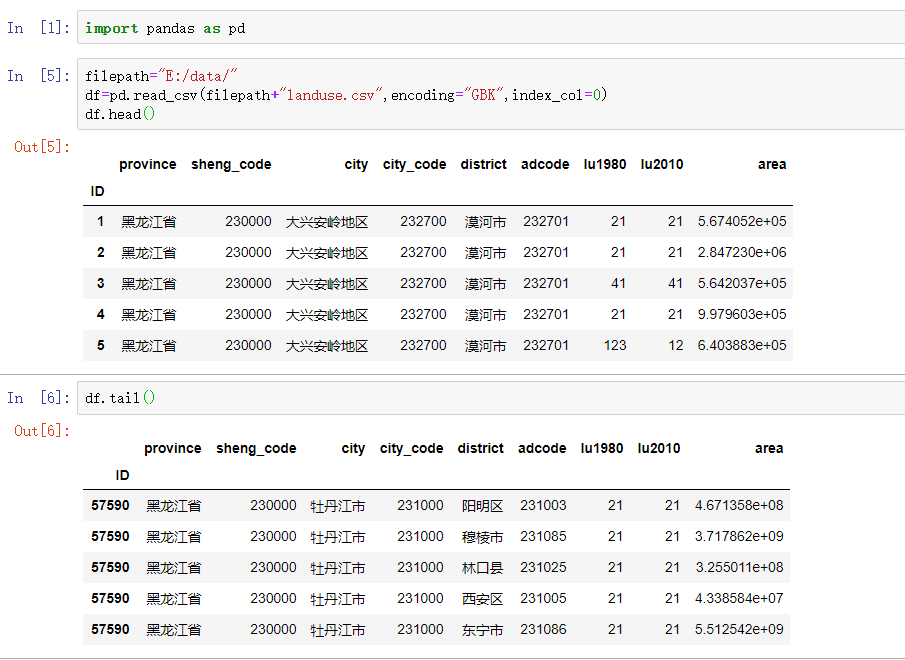
12、将姓名列设置为索引列，然后查询刘梅和王伟光的考试成绩。



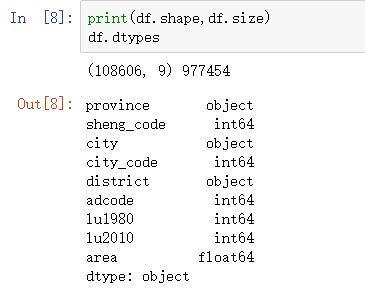
（二）利用pandas相关函数对土地利用数据集进行数据分析

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。

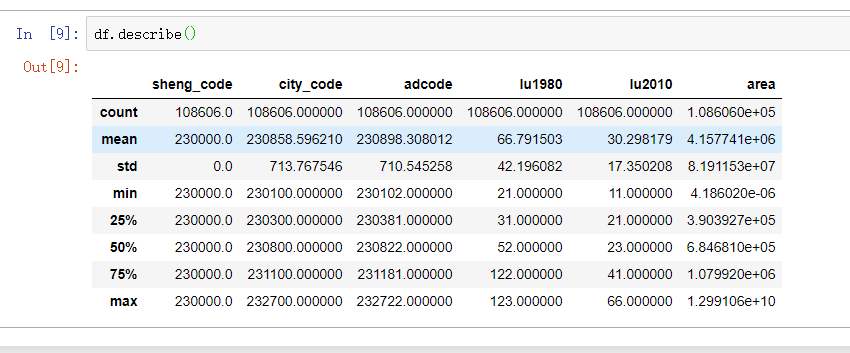
提示：利用pandas的read\_csv函数可以读取csv文件数据



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。



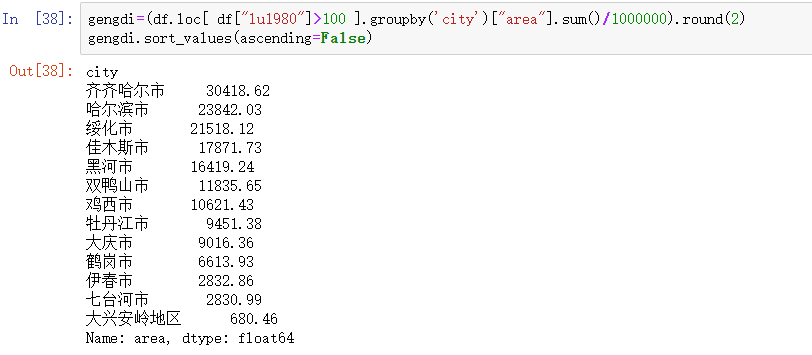
4、查看主要有哪些地区，哪些县市；

提示：利用DataFrame的drop\_duplicates方法可以去除重复值，为了查看方便最好先对数据进行排序，按照所在adcode列根据各县市的行政编码进行排序。

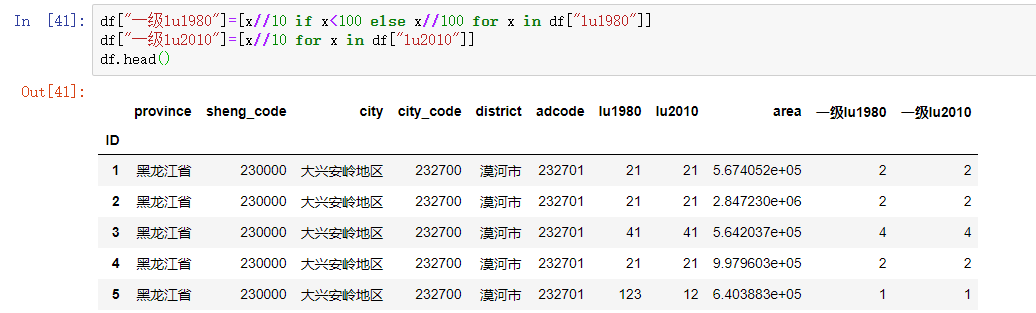




5、查看各地区1980年的耕地面积，要求结果以耕地面积降序排列，以平方公里为单位，保留小数点后两位数字。提示：可以先选择出1980年地类编码>100的记录，然后根据地区名称进行分组后，求取各地区面积的总和，然后按照面积从大到小的顺序进行排序后输出。



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。

提示：

（1）透视表(pivot table)是各种电子表格程序和其他数据分析软件中一种常见的数据汇总工具。它根据一个或多个键对数据进行聚合，并根据行和列上得分组建将数据分配到各个矩形区域中。

透视表的创建可以利用pandas的pivot\_table函数来实现：

pandas.pivot\_table(data,values=None,index=None,columns=None,aggfunc='mean',fill\_value=None,margins=False,dropna=True,margins\_name='All',observed=False)

其中，data为数据对象，values表示需要汇总计算的列；index表示行分组键，一般是用于分组的列名或其他分组键，作为结果DataFrame的行索引；columns：列分组键，一般是用于分组的列名或其他分组键，作为结果DataFrame的列索引；aggfunc：聚合函数或函数列表，默认为平均值。Margins表示是否添加行列的总计，margins\_name表示总计行、列的名称，默认为All。

（2）交叉表实现

交叉表是用于统计分组频率的特殊透视表。简单来说，就是将两个或者多个列重中不重复的元素组成一个新的 DataFrame，新数据的行和列交叉的部分值为其组合在原数据中的数量。

Pandas的crosstab()方法能够快速构建交叉表，并可以通过参数加以个性化的设置。函数原型为：

pd.crosstab(index,columns,values=None,rownames=None,colnames=None,aggfunc=None,margins=False,margins\_name:str='All',dropna:bool=True,normalize=False)

index：类数组，在行中按分组的值。

columns：类数组的值，用于在列中进行分组。

values：类数组的，可选的，要根据因素汇总的值数组。

aggfunc：函数，可选，如果未传递任何值数组，则计算频率表。

rownames：序列，默认为None，必须与传递的行数组数匹配。

colnames：序列，默认值为None，必须与传递的列数组数匹配。

margins：布尔值，默认为False，添加行/列边距（小计）

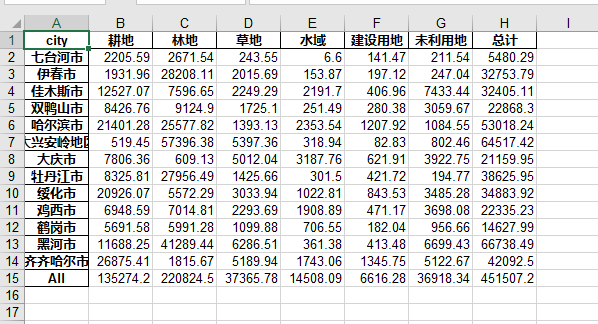
（1）透视表实现方式：



（2）交叉表实现方式：



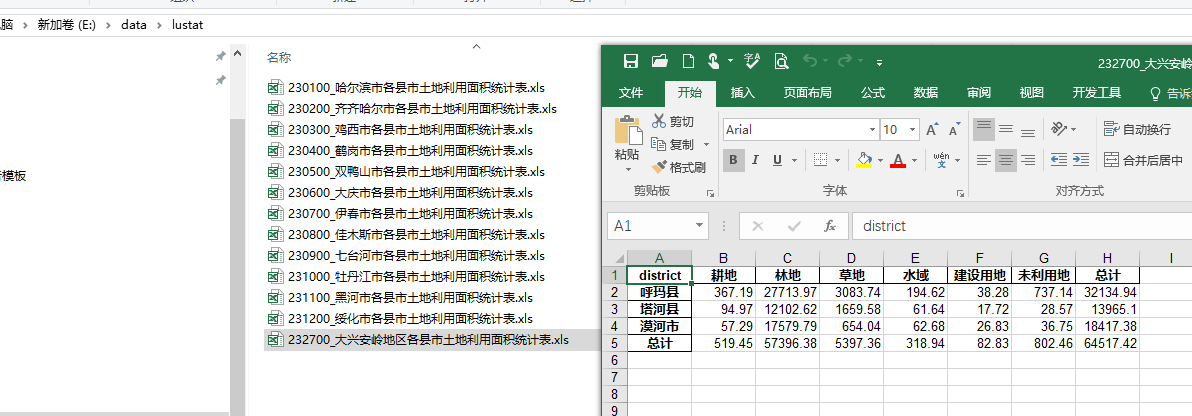
8、将统计的结果保存到excel文件中,文件名为“各县市土地利用面积统计表”。



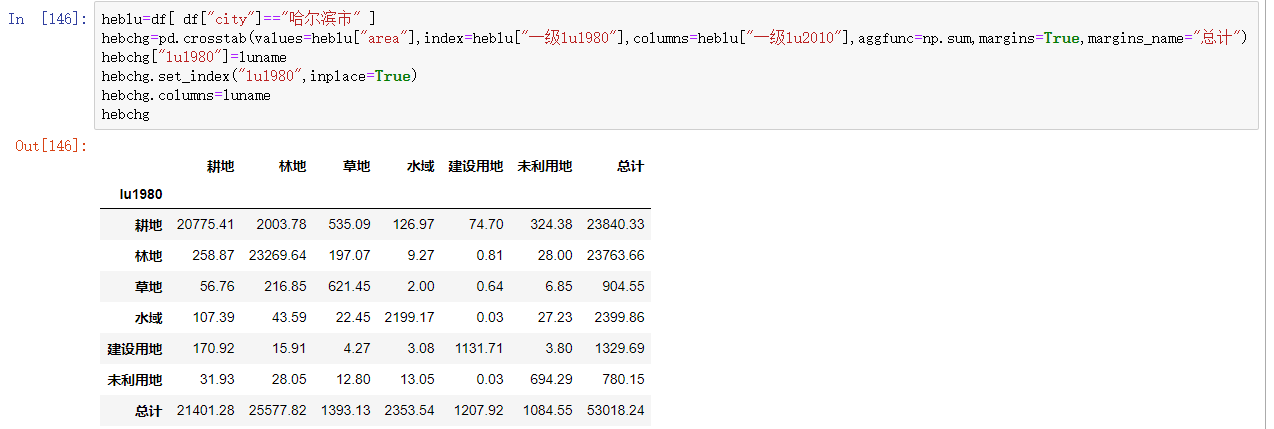
9、将各地区内部不同显示的土地利用数据分别统计，保存到以不同地区命名的文件中，如哈尔滨市保存到stat文件夹下“哈尔滨市各县市土地利用面积统计表.xls”。

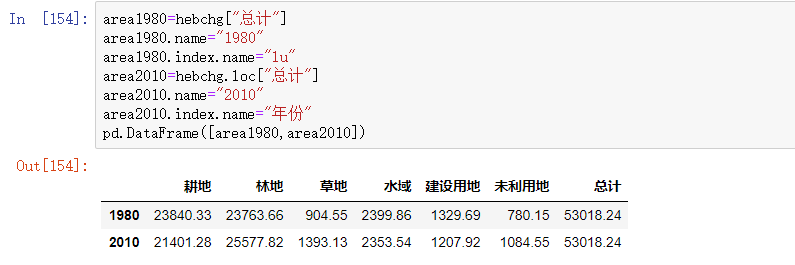


执行结果为：



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。

12、选做题：针对某一类用地的二类用地面积，展开研究。比如研究大兴安岭地区林地中有林地、灌木林地等不同林地类型的面积变化情况，并进行统计与分析。

**五、实验结果与分析：**

（1）验证实现的程序功能与题目要求是否一致。

（2）查找程序出错的语句代码，调试分析出错的原因并改进语句。