**实验五 基于Numpy的数据处理**

**一、实验目的：**

（1）了解Numpy库的基本功能，包括数组的创建、切片、修改元素、排序等方法；

（2）掌握利用Numpy库中的通用函数对数组进行操作与运算的方法；

（3）掌握利用Numpy库对数据进行统计与分析的方法。

**二、实验仪器及实验环境**

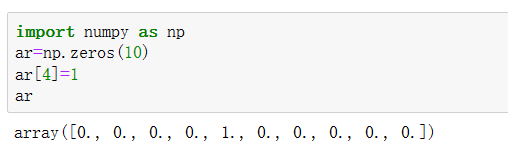
（1）硬件：PC机；

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

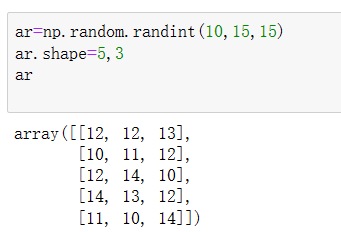
**三、实验内容：**

（一）数组的创建与基本操作

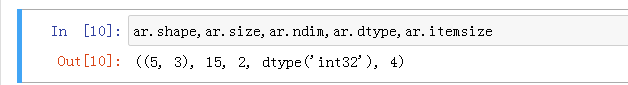
1、创建一个长度为10的一维全为0的ndarray数组，然后让第5个元素等于1。



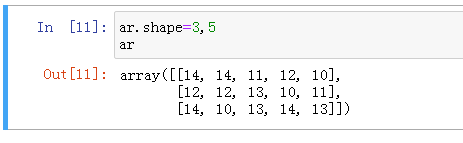
2、创建一个5行3列的二维数组ar，元素大小为从[10-15)区间中随机数组成的整数。



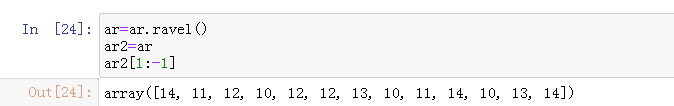
3、查看2题中创建的数组ar的大小、元素个数、维度个数和数组元素的数据类型及数组元素的字节数。



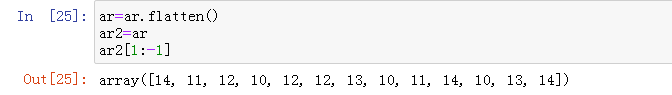
4、将2题中创建的数组修改为3行5列后输出最后一行的元素。



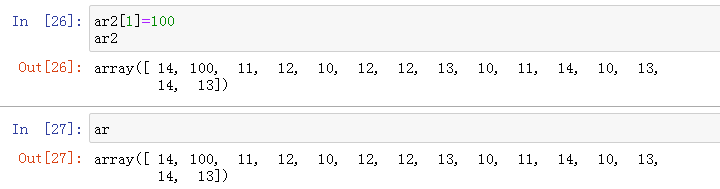
5、将数组ar展平为一维数组，存储至ar2中，然后利用切片提取数组中从下标为1到倒数第二个元素之间的所有元素进行输出；



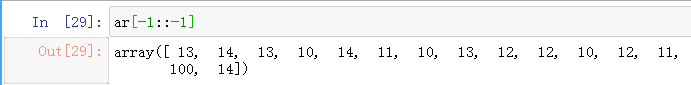
或者：



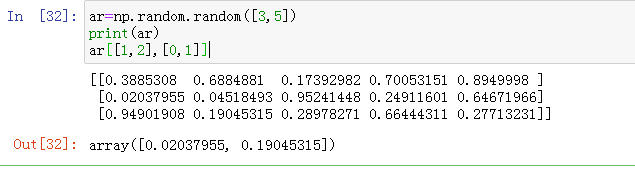
6、将ar2中下标为1的元素修改为100，然后输出ar和ar2数组中各元素的值，分析Numpy中数组切片与Python中列表切片的结果有什么区别。



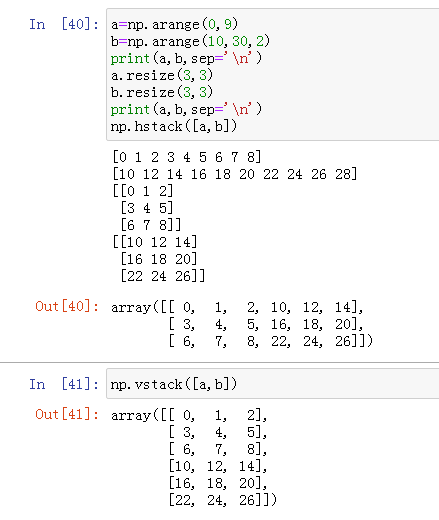
7、对利用切片对ar2中的元素进行反序处理后输出。



8、创建一个二维数组，使用索引方式获取第二行第一列和第三行第二列的元素。

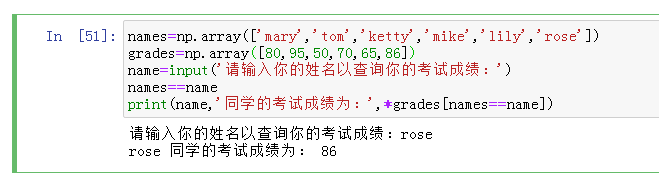


1. 利用创建一个由从0开始到9范围间隔为1的整数组成的数组a，和一个从10到30范围内偶数组成的数组b，然后将a和b转换为3\*3的数组后做水平合并和垂直合并，观察合并后生成的数组元素，比较hstack和vstack函数的功能有什么不同。

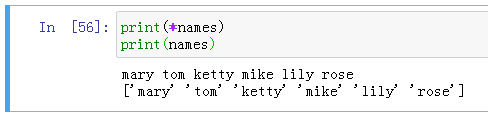


10、创建一个存储6个同学姓名的数组names，和一个存放6名同学考试成绩的数组grades，保证每个同学的姓名和成绩的序号一致，接收用户输入的自己姓名，然后查询该同学的考试成绩。

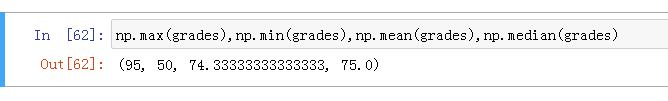
提示：接收到同学的姓名name后，可以使用names==name关系表达式获取该同学的索引，然后用该索引在grades数组中查询相应位置的同学成绩即可。



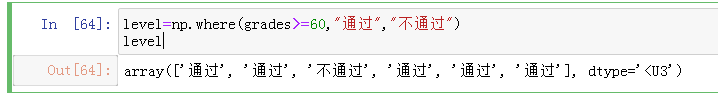
注意：在输出查询结果时，grades[names==name]的结果也是一个数组，输出时应该把数组的边界符 [ ] 去掉，因此需要利用\*来对数组进行解包。其差别可以参看下图：



11、统计第10题中grades数组中的最大值、最小值、平均值、中位数信息，查看学生成绩的统计情况。



1. 根据第10题中的grades数组中的百分制成绩生成二级制成绩单，两级分别为通过和不通过。



（二）利用numpy相关函数对鸢尾花数据集进行数据分析

Numpy是高性能科学计算和数据分析的基础包，里面包含了许多对数组进行快速运算的标准数学函数，掌握这些方法，能摆脱数据处理时的循环。

1、导入isrs.npy文件中的数据

提示：利用numpy的load函数可以读取npy文件数据；

x=np.load('f:\data\isrs.npy',allow\_pickle=True)

2、输出数据集中的数据，查看数据的维度、元素个数；

print(x)

print(x.ndim,x.size,x.shape)

3、查看鸢尾花的类型有哪些，注意去除重复值；

print( np.unique( x[:,-1] ) )

4、获取x中的前4列数据存入data数组，获取最后一列存入flowers数组

data,flowers=x[:,:-1],x[:,-1]

print(data)

5、查看鸢尾花数据集中各属性数据的最大值、最小值、平均值等信息；

np.max(data,axis=0)

6、查看Setosa鸢尾花、Versicolor鸢尾花和Virginica鸢尾花各类样本的数量

print('Setosa',np.sum(flowers=='Iris-setosa'))

print('Versicolor',np.sum(flowers=='Iris-versicolor'))

print('Virginica',np.sum(flowers=='Iris-virginica') )

7、查看Setosa鸢尾花、Versicolour鸢尾花和Virginica鸢尾花花萼的最大值、最小值、均值等数据，分析不同类型鸢尾花数据的特征，选择区分度较大的属性作为后续工作的依据。

for i in np.unique(flowers):

print(i)

datai=data[flowers==i]

print("{:^7s}".format('max'),np.max(datai,axis=0))

print("{:^7s}".format('min'),np.min(datai,axis=0))

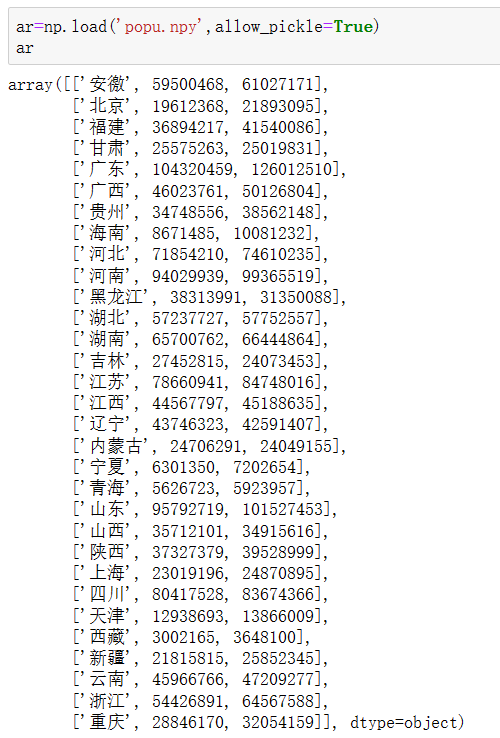
print("{:^7s}".format('median'),np.median(datai,axis=0))

（二）利用numpy相关函数对我国六普和七普人口数据进行数据分析

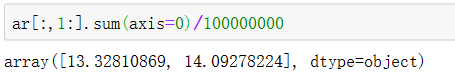
六普数据下载地址：http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexce.htm

1、导入人口数据，数据包中的popu.npy文件是各省六普和七普人口数据

2、显示人口数据，其中第一列是行政区数据，第二列为2010年人口数据，第三列为2020年人口数据

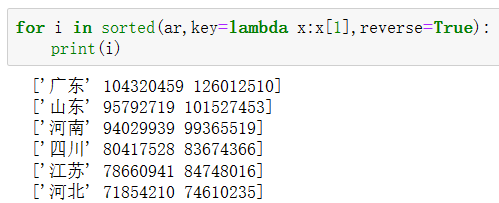


3、统计我国2010年和2020年的总人口

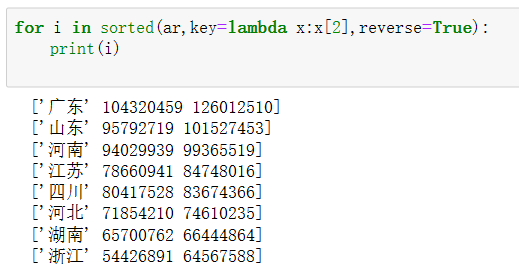


4、将各省数据按照2010年数据降序排列。

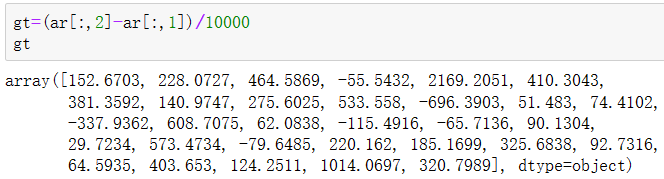
提示：利用python的sorted函数进行灵活的排序方法，需要采用第1列数据进行排序，可以利用lambda表达式编写一个匿名函数来实现提取元素排序键值的功能。查看人数最多的5个省份和人数最少的5个省份。



5、将各省数据按照2020年数据降序排列

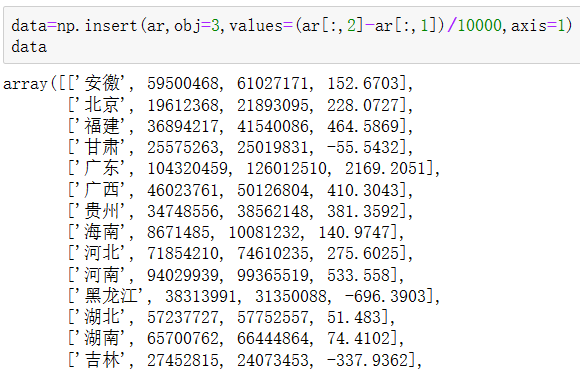


6、计算各省从2010到2020年的人口变化量，单位为万。



1. 将gt中的人口变化数据与ar数组进行横向合并。

提示：gt数组为一维数组，在进行合并前，需要将其转换为二维数组，行数与ar数组相同，列数为1，然后才能与ar数据进行横向合并。



同时，还可以使用numpy的insert函数向数组中添加行（列），使用delete函数删除数组的行或列。

（1）数组行或列的插入函数

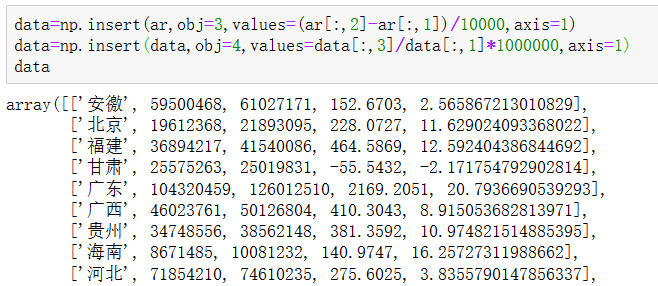
函数原型为：

insert(arr, obj, values, axis=None)

其中，arr为数组对象，obj为需要插入值处的下标切片、下标标量、下标列表、下标整数数组，表示插入的位置，values,为需要插入的集合对象，axis=0表示插入行，1表示插入列。

因此，下面语句可以完成6、7两题中的要求。

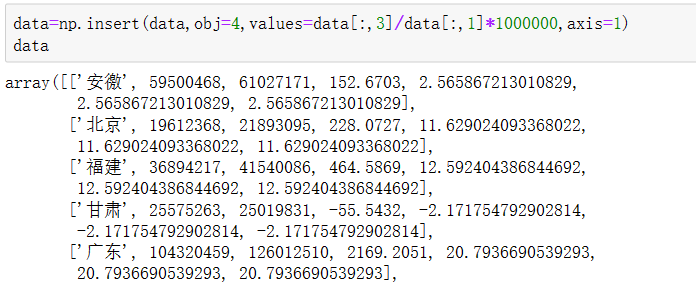
data=np.insert(ar,obj=3,values=(ar[:,2]-ar[:,1])/10000,axis=1)#axis=表示插入列



8、按照同样的方法计算各省份人口变化占2010年总人口的比例，并将该数组作为data的第4列插入到数组中。

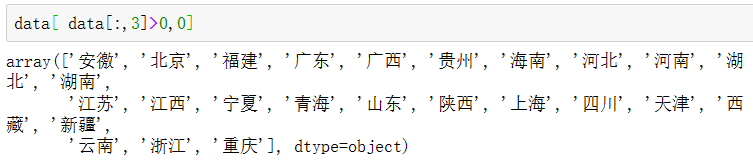
data=np.insert(data,obj=4,values=data[:,3]/data[:,1]\*1000000,axis=1) #表示百分比

完成操作后数组data数据如下图所示。



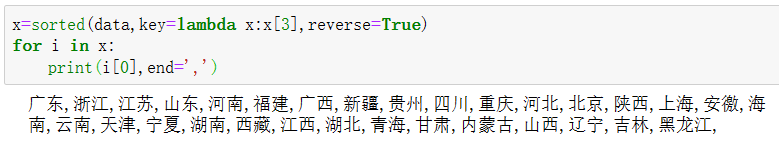
9、查看2010-2020年期间人口增加的省份和人口减少的省份。

data[ data[:,3]>0,0] #利用第3列上值大于0作为行筛选条件

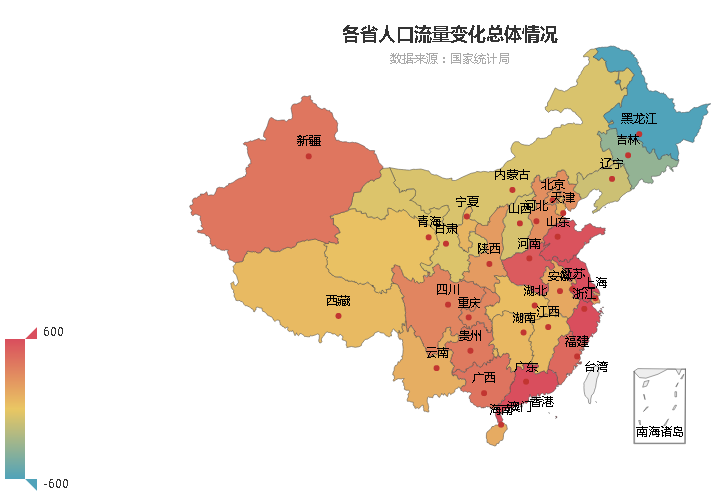


从中可以看出，人口净流入的城市主要为东北、西北和西部城市，中部城市仅有山西省一个省份。

10、将省份按照人口增长的幅度进行降序排列，分析2010-2020年期间各省份人口增长与区域经济、区位等因素的关系。



11、(选做）将结果用地图显示



提示：本题需要先安装pyecharts

pip install pyecharts

data=np.load(r'F:/data/popu.npy',allow\_pickle=True)

data=np.insert(data,obj=3,values=(data[:,2]-data[:,1])/10000,axis=1)#axis=表示插入列

data=data[:,[0,3]]

\_max =600

\_min =-600

map\_ = Map()

map\_.add("", data, maptype="china", zoom=1)

map\_.set\_global\_opts(

title\_opts=opts.TitleOpts(title="各省人口流量变化总体情况",

subtitle="数据来源：国家统计局",

pos\_right="center",

pos\_top="5%"),

visualmap\_opts=opts.VisualMapOpts(max\_=\_max,

min\_=\_min

)

)

map\_.render("F:/data/各省人口流量变化总体情况.html")

**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。