



**《数据分析》实验报告**

**学 院**：

**班 级**：

**学 号**：

**姓 名**：

**指导教师**：

教 务 处

2022 年 4 月

题目

**1. 主成分分析**

**农业生态经济系统各区域单元的有关数据(x1—x9指标，21个样本)如下表1所示，请根据数据完成如下的数据分析任务，包括：**

1. **写入数据库文件，读取数据库文件数据 (略)**

**(2) 对获取得到的数据进行分析**

**1). 相关系数矩阵**

**2). 由相关系数矩阵计算特征值，****以及各个主成分的贡献率与累计贡献率**

**3). 按不低于85%的累计贡献率保留k个主成分，k=\_\_\_\_? 并给出k个主成分的在x1—x9指标下的表达式。**

**4). 如果认为第一主成分z1代表生态经济结构因素，请按照生态经济结构因素对**

**21个区域单元的样本进行由高到低的排名。**



1. **.定义评价系数： ，请按照综合评价**

**系数对样本进行排名。**

**表1 农业生态经济系统各区域单元的有关数据**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 样本序号 | x1：人口密度(人/km2) | x 2：人均耕地面积(ha) | x 3：森林覆盖率(%) | x 4：农民人均纯收入(元/人) | x 5：人均粮食产量 (kg/人) | x 6：经济作物占农作物播面比例(％) | x 7：耕地占土地面积比率(％) | x 8：果园与林地面积之比(％) | x 9：灌溉田占耕地面积之比(％) |
| 1 | 363.912 | 0.352 | 16.101 | 192.11 | 295.34 | 26.724 | 18.492 | 2.231 | 26.262 |
| 2 | 141.503 | 1.684 | 24.301 | 1752.35 | 452.26 | 32.314 | 14.464 | 1.455 | 27.066 |
| 3 | 100.695 | 1.067 | 65.601 | 1181.54 | 270.12 | 18.266 | 0.162 | 7.474 | 12.489 |
| 4 | 143.739 | 1.336 | 33.205 | 1436.12 | 354.26 | 17.486 | 11.805 | 1.892 | 17.534 |
| 5 | 131.412 | 1.623 | 16.607 | 1405.09 | 586.59 | 40.683 | 14.401 | 0.303 | 22.932 |
| 6 | 68.337 | 2.032 | 76.204 | 1540.29 | 216.39 | 8.128 | 4.065 | 0.011 | 4.861 |
| 7 | 95.416 | 0.801 | 71.106 | 926.35 | 291.52 | 8.135 | 4.063 | 0.012 | 4.862 |
| 8 | 62.901 | 1.652 | 73.307 | 1501.24 | 225.25 | 18.352 | 2.645 | 0.034 | 3.201 |
| 9 | 86.624 | 0.841 | 68.904 | 897.36 | 196.37 | 16.861 | 5.176 | 0.055 | 6.167 |
| 10 | 91.394 | 0.812 | 66.502 | 911.24 | 226.51 | 18.279 | 5.643 | 0.076 | 4.477 |
| 11 | 76.912 | 0.858 | 50.302 | 103.52 | 217.09 | 19.793 | 4.881 | 0.001 | 6.165 |
| 12 | 51.274 | 1.041 | 64.609 | 968.33 | 181.38 | 4.005 | 4.066 | 0.015 | 5.402 |
| 13 | 68.831 | 0.836 | 62.804 | 957.14 | 194.04 | 9.11 | 4.484 | 0.002 | 5.79 |
| 14 | 77.301 | 0.623 | 60.102 | 824.37 | 188.09 | 19.409 | 5.721 | 5.055 | 8.413 |
| 15 | 76.948 | 1.022 | 68.001 | 1255.42 | 211.55 | 11.102 | 3.133 | 0.01 | 3.425 |
| 16 | 99.265 | 0.654 | 60.702 | 1251.03 | 220.91 | 4.383 | 4.615 | 0.011 | 5.593 |
| 17 | 118.505 | 0.661 | 63.304 | 1246.47 | 242.16 | 10.706 | 6.053 | 0.154 | 8.701 |
| 18 | 141.473 | 0.737 | 54.206 | 814.21 | 193.46 | 11.419 | 6.442 | 0.012 | 12.945 |
| 19 | 137.761 | 0.598 | 55.901 | 1124.05 | 228.44 | 9.521 | 7.881 | 0.069 | 12.654 |
| 20 | 117.612 | 1.245 | 54.503 | 805.67 | 175.23 | 18.106 | 5.789 | 0.048 | 8.461 |
| 21 | 122.781 | 0.731 | 49.102 | 1313.11 | 236.29 | 26.724 | 7.162 | 0.092 | 10.078 |

问题分析：

1. 通过np.corrcoef()计算相关系数矩阵
2. 通过np.linalg.eig()计算矩阵特征向量和特征值以及各个主成分的贡献率与累计贡献率
3. 通过特征值的累计贡献度选出主成分
4. 按不低于85%的累计贡献率保留k个主成分

代码：

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from sklearn import preprocessing

plt.rcParams['font.sans-serif'] = ['SimHei']

plt.rcParams['axes.unicode\_minus'] = False

data1 = pd.read\_csv('2.csv', encoding='gbk')

data1 = data1.set\_index(['样本序号'])

data1

data2 = preprocessing.scale(data1)

data2

# 计算相关系数矩阵

data3 = np.around(np.corrcoef(data2.T),decimals=3)

data3

# 计算特征值，特征向量

featValue, featVec= np.linalg.eig(data3.T)

featValue, featVec

#求特征值的贡献度

gx = featValue/np.sum(featValue)

gx

#求特征值的累计贡献度

lg = np.cumsum(gx)

lg

#选出主成分 k = 3 三个特征值贡献率相加>85

k = [0, 1, 2]

k#选出主成分对应的特征向量矩阵

selectVec = np.matrix(featVec.T[k]).T

selectVe=selectVec\*(-1)

selectVec

df1=data1.values\*np.array(pd.DataFrame(selectVec).loc[:,0])

tf1=df1\*gx[0]/100

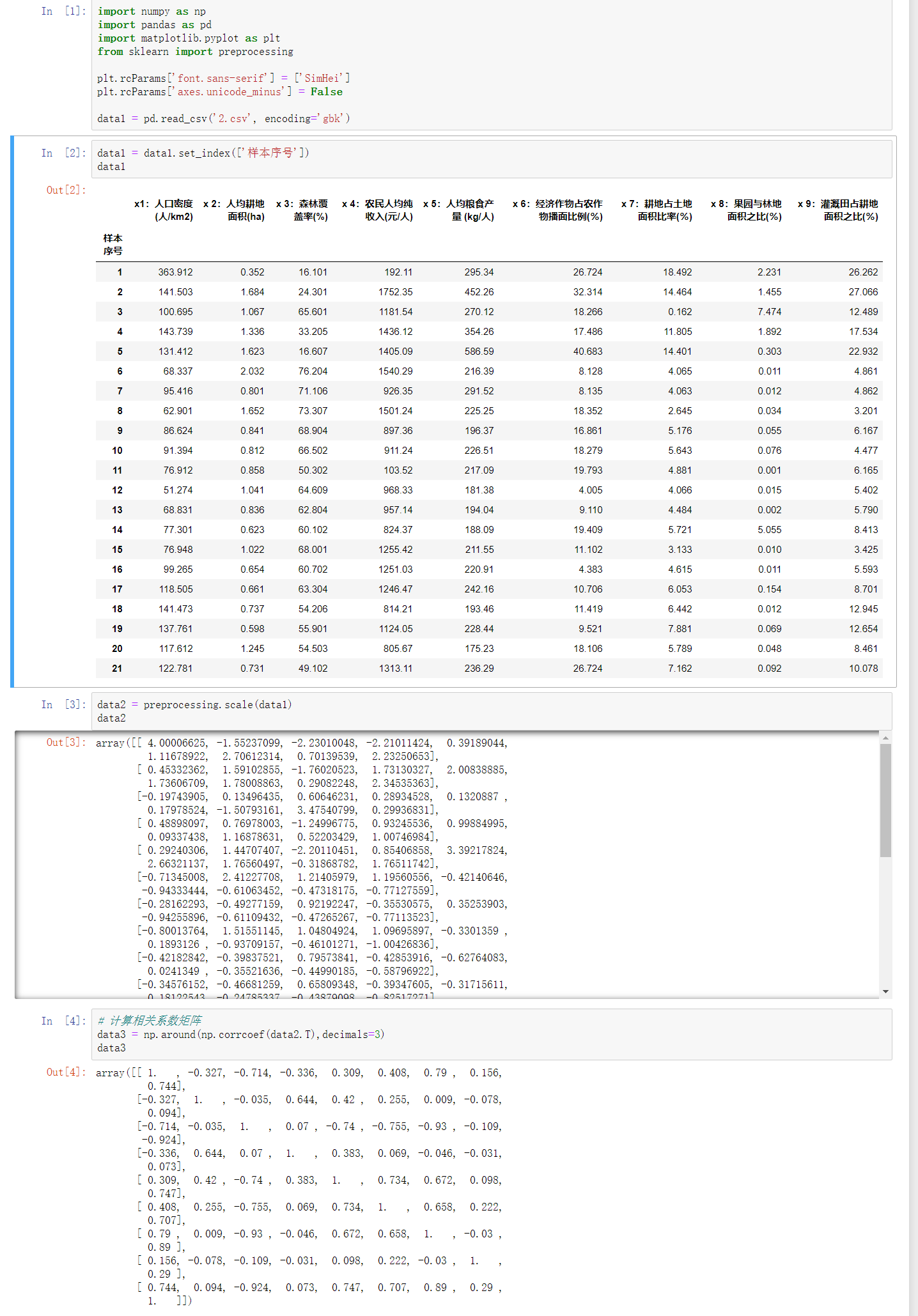
pd.DataFrame(tf1,index=data1.index, columns=data1.columns).sort\_values(["x1：人口密度(人/km2)"])

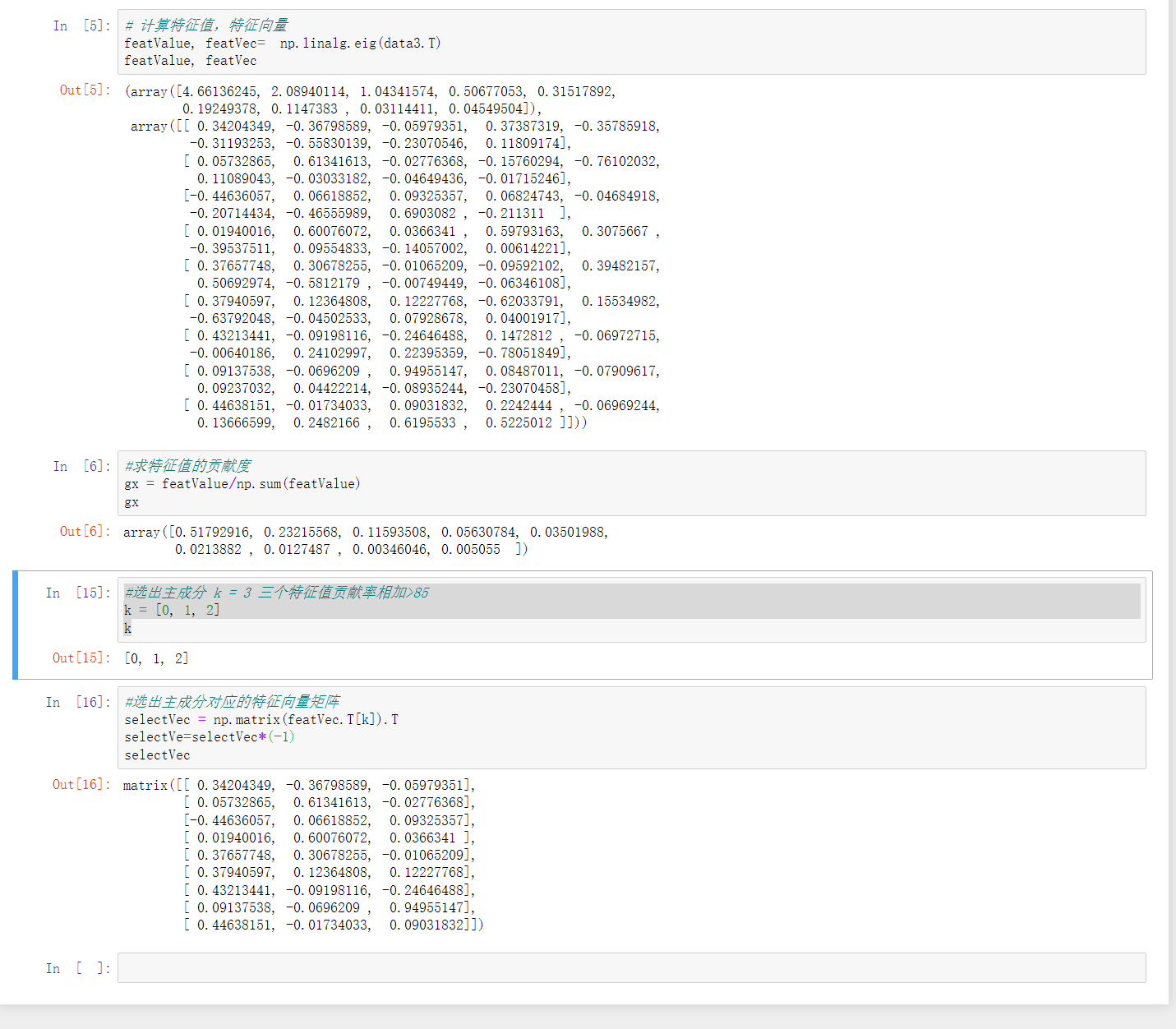
df=data1.values\*np.array(pd.DataFrame(selectVec).loc[:,0])+data1.values\*np.array(pd.DataFrame(selectVec).loc[:,1])+data1.values\*np.array(pd.DataFrame(selectVec).loc[:,2])

tf=df\*gx/100;

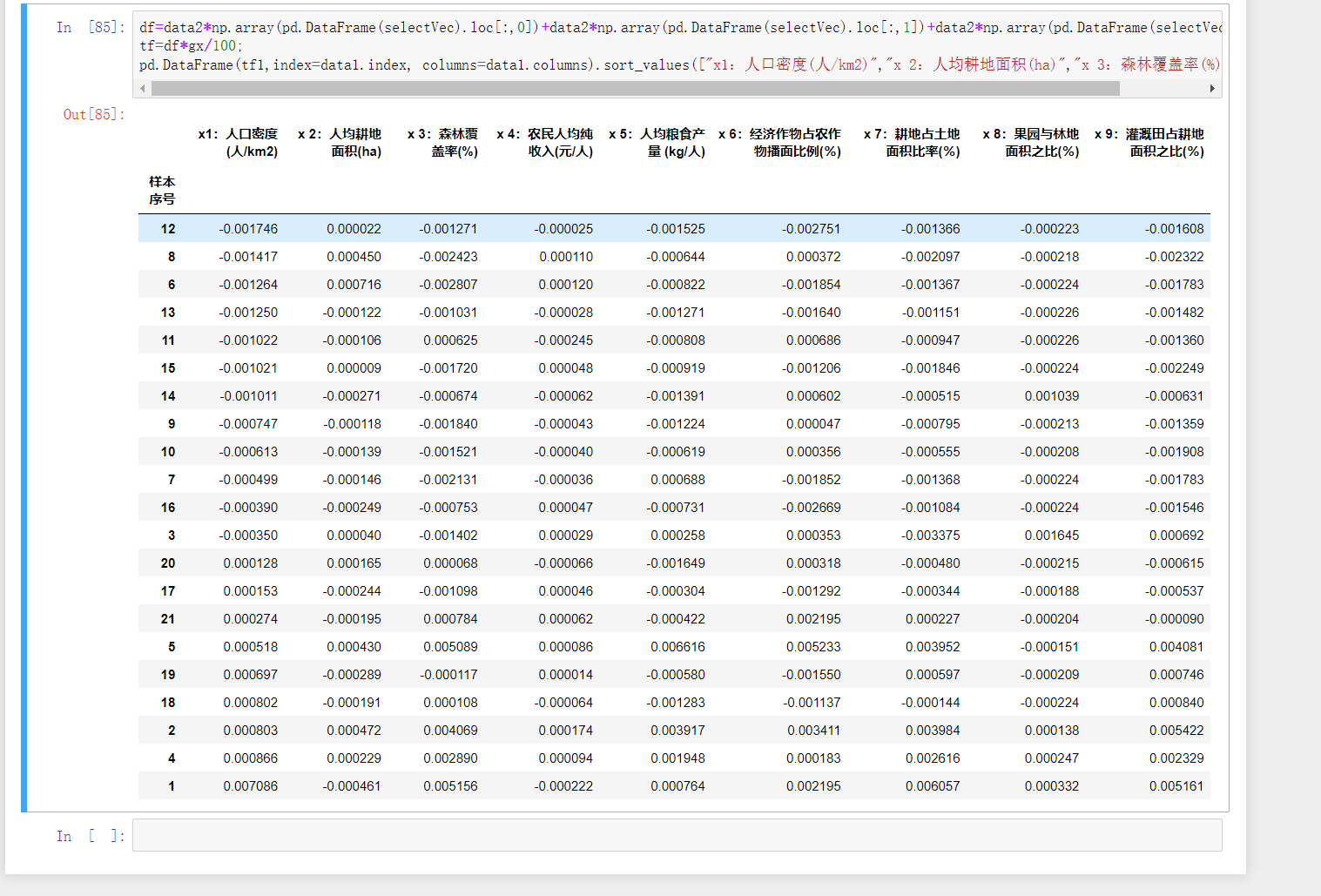
pd.DataFrame(tf1,index=data1.index, columns=data1.columns).sort\_values(["x1：人口密度(人/km2)","x 2：人均耕地面积(ha)","x 3：森林覆盖率(%)"])

结果与分析：









实验过程记录：

较为顺畅，在计算特征值之前进行了标准化。