|  |  |  |
| --- | --- | --- |
| **得分** | **教师签名** | **批改日期** |
|  |  |  |

课程编号： IB01017

****

**深圳技术大学实验报告**

**课程名称： Python 语言程序设计**

**实验名称： Numpy函数包的使用**

**班 级： 计算机科学与技术2班**

**指导教师： 柯笑**

**报 告 人： 黄荣权 学号： 202002020213**

**合 作 者： 组号：**

**实验地点： C1-405**

**实验时间： 2022 年 06 月 13 日 星期 一**

**提交时间： 2022/06/13**

## 实验十六 Numpy函数包的使用

1. 实验学时

2学时

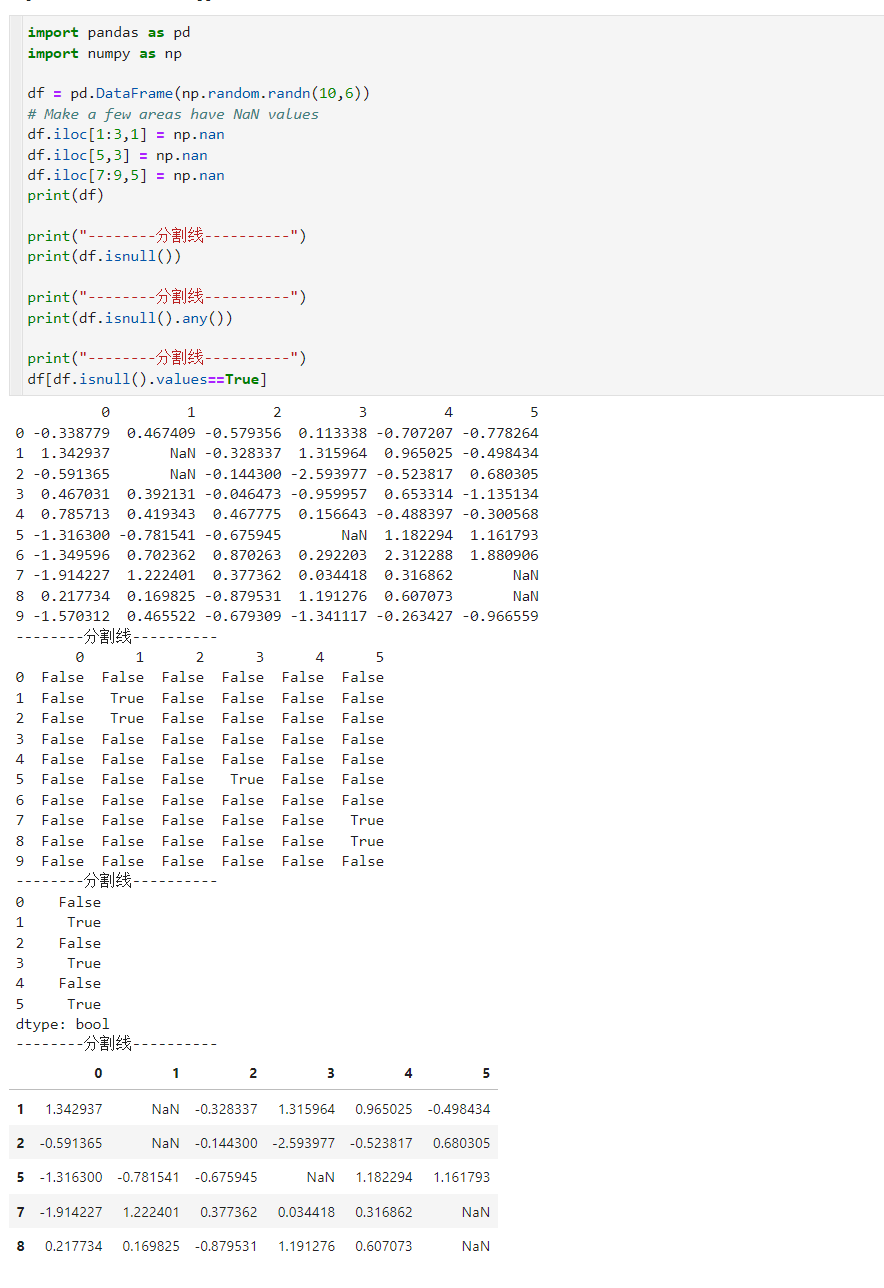
### 实验目的

### 熟悉Numpy库

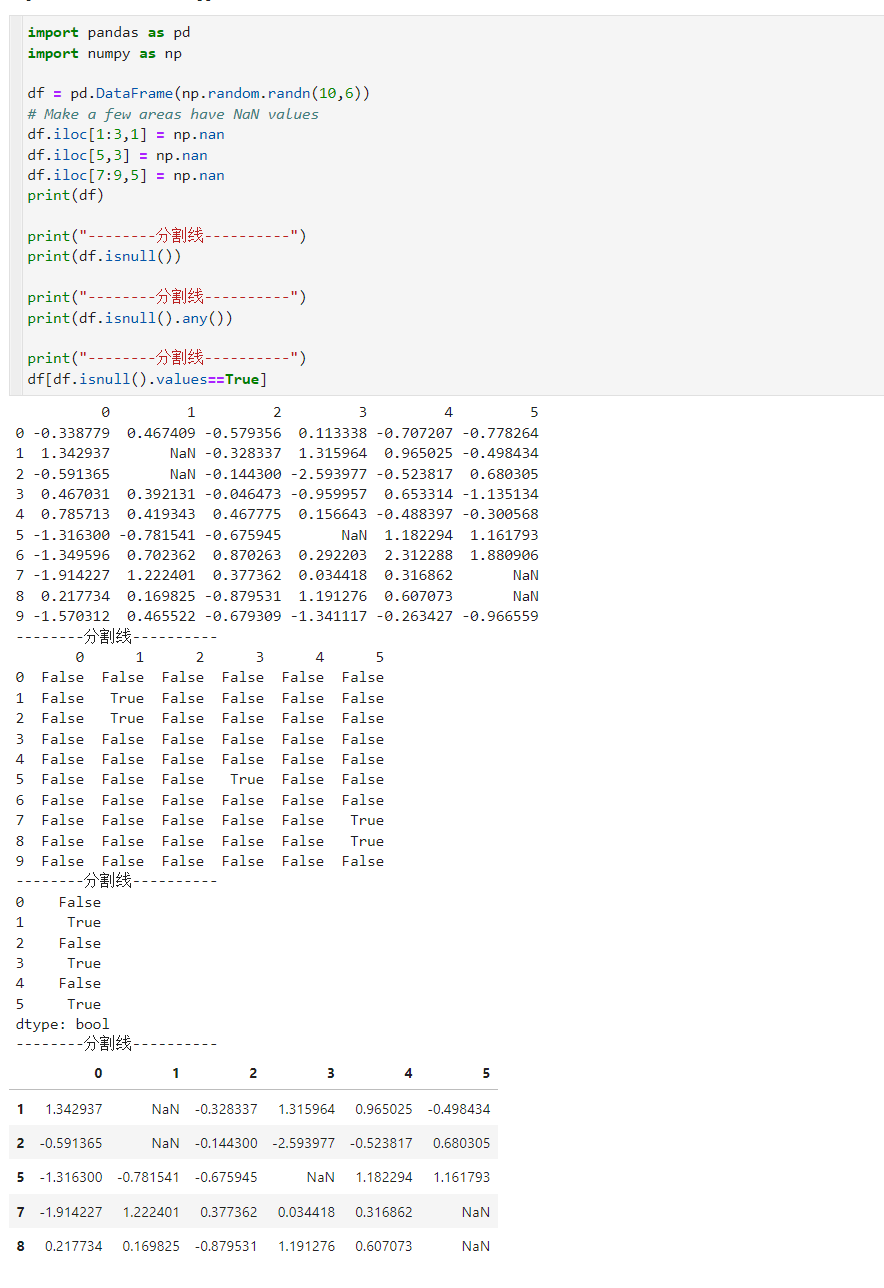
### 掌握缺失值查找的方法以及相关的处理。

### 三、实验内容

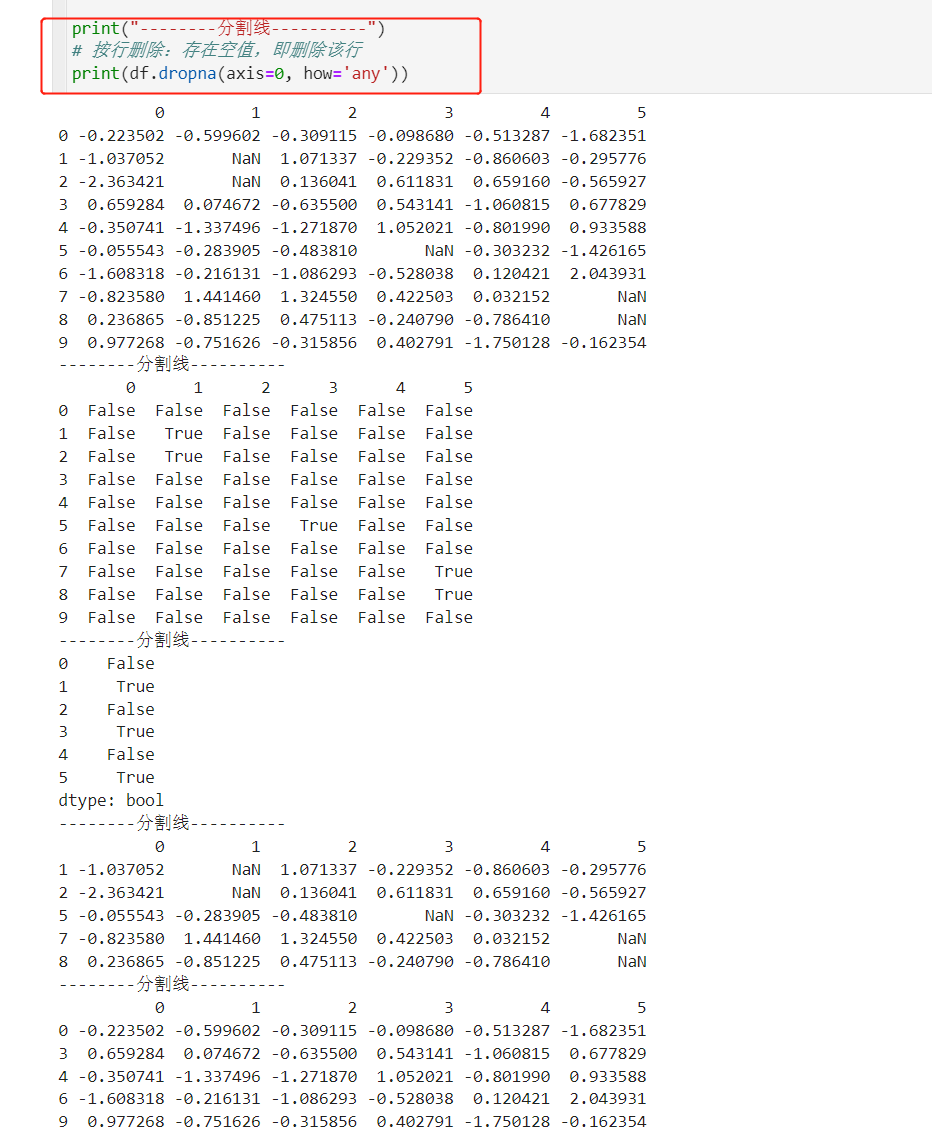
### （一）利用np.random.rand函数创建一个数组。并随机的在这个数组中添加缺失值。



### 统计出出现缺失值的总数以及返回缺失值的位置。



### 删除缺失值所在的行。



### 四、思考题

（一）缺失值处理有哪些方法？

四种方法：

(1)均值插补。数据的属性分为定距型和非定距型。如果缺失值是定距型的，就以该属性存在值的平均值来插补缺失的值；如果缺失值是非定距型的，就根据统计学中的众数原理，用该属性的众数(即出现频率最高的值)来补齐缺失的值。

(2)利用同类均值插补。同均值插补的方法都属于单值插补，不同的是，它用层次聚类模型预测缺失变量的类型，再以该类型的均值插补。假设X=(X1,X2…Xp)为信息完全的变量，Y为存在缺失值的变量，那么首先对X或其子集行聚类，然后按缺失个案所属类来插补不同类的均值。如果在以后统计分析中还需以引入的解释变量和Y做分析，那么这种插补方法将在模型中引入自相关，给分析造成障碍。

(3)极大似然估计（Max Likelihood ,ML）。在缺失类型为随机缺失的条件下，假设模型对于完整的样本是正确的，那么通过观测数据的边际分布可以对未知参数进行极大似然估计（Little and Rubin）。这种方法也被称为忽略缺失值的极大似然估计，对于极大似然的参数估计实际中常采用的计算方法是期望值最大化(Expectation Maximization，EM）。该方法比删除个案和单值插补更有吸引力，它一个重要前提：适用于大样本。有效样本的数量足够以保证ML估计值是渐近无偏的并服从正态分布。但是这种方法可能会陷入局部极值，收敛速度也不是很快，并且计算很复杂。

(4)多重插补（Multiple Imputation，MI）。多值插补的思想来源于贝叶斯估计，认为待插补的值是随机的，它的值来自于已观测到的值。具体实践上通常是估计出待插补的值，然后再加上不同的噪声，形成多组可选插补值。根据某种选择依据，选取最合适的插补值。

多重插补方法分为三个步骤：①为每个空值产生一套可能的插补值，这些值反映了无响应模型的不确定性；每个值都可以被用来插补数据集中的缺失值，产生若干个完整数据集合。②每个插补数据集合都用针对完整数据集的统计方法进行统计分析。③对来自各个插补数据集的结果，根据评分函数进行选择，产生最终的插补值。

# 五、实验结论或体会

通过此次实验，我熟悉了Numpy库，掌握了缺失值查找的方法以及相关的处理。

|  |
| --- |
| **指导教师批阅意见：** |
| **成绩评定：**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **操作及记录**  （50分） | **实验总结**  （20分） | **思考题**  （10分） | **报告整体印象**  （20分） | **总分** | |  |  |  |  |  | |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。

2、教师批改学生实验报告时间应在学生提交实验报告时间后10日内。