(192)上海大学2019-2020年冬季学期试卷(A)　2020.6

课程名：数据分析与智能计算 课程号：00864118 学分：3

基础题A

应试人声明:

我保证遵守《上海大学学生手册》中的《上海大学考场规则》，如有考试违纪、作弊行为，愿意接受《上海大学学生考试违纪、作弊行为界定及处分规定》的纪律处分。

学号：（见登录信息）　　姓名：（见登录信息）

题目	选择题	填空题				编程题		总分
题号	1~15	1	2	3	4	1	2	
题分	30	5	5	10	20	20	10	100
得分								

本试卷由选择题（30分）、程序填空题（40分）和编程题（30分）三部分组成，

选择题共包括15个单选题，由计算机自动完成组卷和阅卷。

（本试卷考试时间　90　分钟）

一、单选题（本大题 15 道小题，每小题 2 分，共 30 分），从下面题目给出的 A、B、C、D 四个可供选择的答案中选择一个正确答案。

1. 将一维数组转化为多维数组的 numpy 函数是_____。

A. arange()

B. reshape()

C. zeros()

D. ones()

2. DataFrame 对象的列索引通常表示_____。

A. 列的位置信息

B. 每列数据的总数

C. 列的数据类型

D. 每列数据对应的现实概念

3. DataFrame 对象 df 中基于位置序号选取第 2 行第 3 列数据的方式是_____。

（序号从 0 开始）

A. df.rloc[1,2]

B. df.loc[1,2]

C. df.find(1,2)
D. df.iloc[1,2]

4. CSV 文件是_____，可以使用_____查看。
A. 纯文本文件，文本编辑器
B. ppt 文件，powerpoint 查看
C. word 文件，word 查看
D. 图像文件，画图工具查看

5. 关于 DataFrame 和 Series 对象，下列叙述正确的是_____。
A. DataFrame 对象只能用于处理两维数据
B. DataFrame 对象不能转化为 Series 对象
C. Series 对象可以用来处理多维数据
D. Series 对象主要用于处理一维数据

6. 统计量“方差”描述_____。
A. 样本个体距离均值的离散程度
B. 样本中出现次数最多的值
C. 样本（一组数据）的平均值
D. 样本中不同的值占样本容量的比例

7. 假定 DataFrame 对象 temp 中共有 12 列，语句_____删除空值（NaN）个数大于 3 的行。
A. temp.dropna(threshold = 8)
B. temp.dropna(threshold = 10)
C. temp.dropna(threshold = 9)
D. temp.dropna(threshold = 7)

8. names=np.array(['马化腾','李彦宏','雷军','扎克伯格'])，names[2]的值是_____。
A. 马化腾
B. 李彦宏
C. 雷军
D. 扎克伯格

9. 记录同学成绩的 scores 数组如下，scores[1:3, [2,5]] 取得的数据是_____。
scores： array([[70, 85, 77, 90, 82, 84, 89],
[60, 64, 80, 75, 80, 92, 90],
[90, 93, 88, 87, 86, 90, 91],
[80, 82, 91, 88, 83, 86, 80],
[88, 72, 78, 90, 91, 73, 80]])
A. array([[80,92], [88,90]])
B. array([[64,80], [93,86]])
C. array([[85,84],[64,92], [93,90]])
D. array([[64,92], [93,90],[82,86]])

10. 比较 3 个班级学生高数成绩的分位数分布并观察异常值，可选择_____。

A. 直方图

B. 密度图

C. 箱须图

D. 柱状图

11. 绘制多个子图的正确方法是_____。

A. 导入 matplotlib.pyplot 库，创建 figure 对象，调用 figure.subplot 函数

B. 导入 pandas.pyplot 库，创建 figure 对象，调用 figure.subplot 函数

C. 导入 matplotlib.pyplot 库，创建 figure 对象，调用 figure.add_subplot 函数

D. 导入 pandas.pyplot 库，创建 figure 对象，调用 figure.add_subplot 函数

12. subjects=np.array(['Math', 'English', 'Python', 'Chinese', 'Art', 'Database', 'Physics'])，mask=(subjects=='English')|(subjects=='Art')。则 mask 数组中值为 True 的元素个数是_____。

A. 2

B. 3

C. 4

D. 5

13. 使用 merge 方法对 DataFrame 对象 temp1 和 temp2 进行列上的合并时，设置参数_____，实现按照两个对象键值的交集进行合并。

A. how=left

B. how=inner

C. how=right

D. how=outer

14. 下面关于数据科学与大数据之间的关系描述，错误的是_____。

A. 大数据属于数据科学的范畴

B. 大数据分析遵循数据科学处理问题的基本工作流程

C. 大数据分析采用的技术完全不同于数据科学技术

D. 大数据技术是指数据量达到某种规模时引入的分布式存储.计算和传输等方法

15. 下面关于使用 pyplot 和 pandas 提供的绘图函数的说法中，错误的是_____。

A. pandas 提供的绘图函数使用更快捷

B. 相比较 pandas 绘图，pyplot 提供更多图元绘制函数，能提供更精细的绘图方式

C. Series.DataFrame 对象都提供 plot()函数

D. 在同一 figure 对象中，pyplot 和 pandas 的绘图函数不可以混合使用

填空题答案

1. 填充

2. kde

3. fillna　inplace

4.

【1】random.normal

【2】size=(5,8);(5,8)

【3】axis=1

【4】(arr>300)

二、填空题(本大题 4 道小题，每空 5 分，共 40 分)。

1. 清洗数据有滤除和填充两种方法，当数据集比较小时，应尽量选择数据_____的方式来清洗数据。

2. 利用 Series.plot 绘制概率密度图时，要将 kind 参数设置为_____。

3. 调用 DataFrame 对象 temp 的_____方法填充空值时，设置_____参数可以控制是否直接更新 temp 对象。

4. 利用随机函数模拟果汁生产线上每瓶饮料的实际装瓶容量。饮料装瓶核定容量为 300ml，实际装瓶容量的方差为 5ml（服从均值为 300.方差为 5 的正态分布）。假设抽检 5 个批次，每次 8 瓶果汁样品。

1）生成 5*8 的数组保存每瓶的实际容量并显示；

2）输出每个批次装瓶容量的实际均值（输出小数位限制为 2 位）；

3）统计所有抽检样品中装瓶容量大于 300ml 的个数。

源程序代码如下，其中【1】【2】【3】【4】为需要填空的部分

```
import numpy as np
#设置显示精度为两位小数
np.set_printoptions(precision=2, suppress=True)
#按照正态分布随机生成 5*8 的数组模拟装瓶容量，并输出
arr = np.【1】(300,5,【2】)
print ("1. \n", arr)
print("2. \n", arr.mean(【3】) )
#输出每个批次装瓶容量的实际均值
print("3.\n",【4】.sum() )
#统计所有抽检样品中装瓶容量大于 300 的个数。
```

三、简答题(本大题 2 道小题，共 30 分)

1. 葡萄酒数据集（wine.data）搜集了法国不同产区葡萄酒的化学指标。请完成如下分析功能：

1）从数据集文件（注：这是个 csv 文件）中读出数据，保存到 DataFrame 中（3 分）；

2）判断数据集中是否有缺失数据，如有缺失请删除包含缺失数据的行（3 分）；

3）把数据集分为测试集和训练集（4 分）；

4）在训练集上建立决策树模型，并分析此模型在测试集上的效果。

（需要计算准确率.打印分类报告和混淆矩阵并简要说明决策树模型的性能）（10 分）；

请编写程序实现上述功能要求，并将完成的程序（.py 或.ipynb 文件）上传。

2. 简述你对 "数据科学与人工智能"应用与发展的认识与思考，末尾署名（学号+姓名）。