(202)上海大学2020-2021年冬季学期模拟试卷　2021.3

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	程序题			总分
题号	1~15	1	2	3	
题分	30	20	20	30	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. 下列不属于数据分析应用场景的是_____
   A. 产品销量分析
   B. 码头货物吞吐量预测
   C. 计算机硬盘使用寿命预测
   D. 某人一生的命运预测
2. 下列不属于 Python 优势的是_____
   A. 语法简洁，程序开发速度快
   B. 拥有大量的第三方库，能够调用 C、C++、Java 语言
   C. 程序的运行速度在所有计算机语言中最快
   D. 开源免费
3. 下列不属于 Numpy 库的 ndarray 数组属性的是_____
   A. ndim
   B. shape
   C. size
   D. add

4. 如何生成由 5 个随机整数组成的一维数组，整数取值范围为 0-10。
下面选项中正确的是_____
A. np.random.randint(0,11,5)
B. np.random.randint(0,10,5)
C. np.randint(0,11,5)
D. np.randint(0,10,5)
5.下列不能实现将 shape 为 dtype[5,7]的 scores 数组所有元素都加 10 的语句是_____
A. scores + 10
B. np.add(scores, 10)
C. scores[10].add(10)
D. scores + [10,10,10,10,10,10,10]
6. CSV 文件是常用的数据文件格式，可以使用_____查看
A. 文本编辑器、Excel
B. photoshop
C. powerpoint
D. 画图工具
7. 下面关于 DataFrame 存储表结构数据的说法，错误的是_____
A. 通常使用行存储一条数据，列存储该数据的各个特征项
B. DataFrame 对象只能使用行、列索引对进行数据切片，不能使用位置序号
C. 从 DataFrame 对象中取出一列，得到 Series 对象
D. Series 对象可以使用 Numpy 的函数进行统计分析
8. 关于数据文件读写，_____是错误的描述
A. pandas 读取的数据文件中可以包含中文字符组成的数据
B. 文件中第一行必须给出列的索引名（columns），否则 pandas 无法读取各列内容
C. 读取 excel 文件时，可以为 sheetname 参数赋值，以读取指定表单的数据
D. csv 数据文件用换行符来区分数据行
9. Python 的工具包_____设计了两种数据结构 Series 和 DataFrame
A. numpy
B. pandas
C. matplotlib
D. scikit-learn
10. 某人做数据分析测得个人健康和年龄的相关系数是 -1.09。根据这个你可以得出哪个结论_____
A. 年龄是健康程度很好的预测器
B. 年龄是健康程度很糟的预测器
C. 这个相关系数有错误
D. 以上说法都不对
11. 下面哪个函数的作用是显示图形并关闭此次绘图_____
A. plt.figure()
B. plt.show()
C. plt.savefig()
D. plt.plot()

12. _____可用于展示离散数据

A. 折线图

B. 柱状图

C. 统计地图

D. 曲面图

13. 为描述高校教师学历占比情况，适合的图形是_____

A. 散点图

B. 曲面图

C. 直方

D. 饼图

14. F1_score 可用于衡量分类模型性能，根据混淆矩阵，F1 = _____

A. 2a/(2a+b+c)

B. (a+d)/(a+b+c+d)

C. a/(a+c)

D. a/(a+b)

15. 关于聚类分析，正确的是_____

A. "簇"越少说明聚类效果越好

B. 聚类是有监督学习方法

C. 聚类可作为分类等其他任务的预处理过程

D. 同一个数据集，不同的聚类算法得到的结果是一样的

二、编程题

编程题 1（30 分）

表 1 和表 2 给出了四种食物每百克蛋白质、脂肪和碳水化合物的含量数据（克）。请填空完成如下程序功能：

1） 根据表 1 和表 2 中内容分别创建 DataFrame 对象 df1 和 df2，将 df2 中内容与 df1 内容合并，并都存放在 df1 中；

2） 判断 df1 每一列中是否存在空值，并打印输出判断结果（True 或者 False）

3） 删除数据项缺失数量（有空值）大于等于 2 的行；

4） 用列均值填充相应的缺失数据（原位修复）并打印输出修复后数据。

	蛋白质	脂肪	碳水化合物
米饭	2.6	0.3	25
牛肉	20.2	NAN	NAN

	蛋白质	脂肪	碳水化合物
鸡蛋	NAN	10.1	1.4
牛奶	3	3.2	3.4

请填空完成程序，需要填空的程序为素材文件夹【E:\KS\sc182003001】中的 source.py,其中【1】【2】【3】【4】为需要填空的部分,将填空后完整的程序以文件名 182003001.py 保存到考试文件夹【E:\KS\】文件夹下。

"

编程题 2（40 分）

数据集（素材文件夹【E:\KS\sc182005001】中的 DataScience.xls）记录了某系的实验教学计划，请编写程序完成下述功能要求，并将完成的程序（.py 或.ipynb 文件）以文件名 182005001 保存到考试文件夹【E:\KS\】文件夹下

（1） 读取 DataScience.xls 文件数据，创建为 data 数据对象

（2） 查询 df 的数据量和基本结构（df.index，df.columns）

（3） 查询 df 中是否含有 NaN 数据？将含有 NaN 数据的行导出为数据文件 pre.csv，判断采用何种数据清洗模式：填充、删除或手工填充

（4） 查询课程名称、实验项目名称、实验类型和二级实验室四列数据内容

（5） 统计每一门课程的实验课时数

（6） 统计每周开设所有实验课时数

（7） 统计每门课程的实验类型分布

（8） 统计每个班级的实验课课表

（9） 分析各二级实验室承担的实验课时量

（10） 分析各二级实验室能够支持的实验类型