2024 年上海市高等学校信息技术水平考试试卷

二三级 数据科学技术及应用（A 场）

（本试卷考试时间 150 分钟）

一、单选题（ 本大题 12 道小题 ，每小题 2 分，共 24 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.无人机通过搭载的传感器搜集农场周边的环境数据，这种数据获取方法是_____。

A.人工采集

B.设备采集

C.系统日志采集

D.网络爬虫采集

2.在举办城市马拉松比赛时，组委会通常按固定距离设置沿途服务站点，确定服务站点数量后，可以使用_____函数方便地计算站点位置（距起点距离）的序列。

A.numpy.linspace

B.numpy.arange

C.numpy.random.randint

D.numpy.random.normal

3.下面语句建立一个水果含糖量(单位：克/百克）的数据对象：fruit = pd.DataFrame([4.0,10.7,16,5.7],columns=["含糖量"],index=["西瓜","桃子","荔枝","草莓"])，下列选项中能读取荔枝含糖量的是_____。

A.fruit["荔枝"]

B.fruit.loc["荔枝"]

C.fruit.loc[name="荔枝"]

D.fruit.荔枝

4.关于 DataFrame 对象的 dropna()函数，在使用默认参数值时，说法错误的是_____。

A.删除空值所在的行或列

B.删除后产生新数据对象

C.原始对象的数据也会删除

D.参数 axis 为 0 表示按行删除

5.数组 CAP 记录了某品牌果汁的实际装瓶容量，可以使用_____统计量来分析实装容量的数据离散程度。

A.均值

B.方差

C.中位数

D.众数

6.食堂收集了学生近 10 年的平均用餐费用，使用_____适合分析展示就餐费用变化趋势。

A.直方图

B.饼图
C.折线图
D.箱须图

7.张皓云一家的身高如下：爸爸 1.7 米，妈妈 1.54 米，叔叔 1.69 米，李青 1.78 米，李青儿子 1.71 米。李青家的身高可以用_____的理论解释。
A.回归分析
B.聚类分析
C.分类分析
D.神经网络

8.下列应用场景，采用无监督学习方法的是_____。
A.标注花草图像的类别，训练模型识别花草类型
B.使用 PCA（主成分分析）实现高维数据集的特征筛选
C.AI 围棋软件学习围棋古谱，提高对弈水平
D.使用历史有标记的网站访问行为数据建立异常检测模型

9.有监督分析建模时通常将数据集分为训练集和测试集，以下说法错误的是_____。
A.通常训练集的样本数量大于测试集
B.测试集的数据应从原始数据集中随机进行选取
C.训练后的分析模型，通常在测试集上的性能会降低
D.模型训练时，训练集和测试集一起参与训练，性能评估时只采用测试集

10.在使用手机导航时，司机可以通过说出指令来控制导航 APP 显示路径，这项技术主要采用了_____方法。
A.图像识别
B.语音识别
C.聚类技术
D.降维技术

11.在教学评价软件中，利用_____可以分析学生评语的情感倾向性。
A.图像识别技术
B.文本分类技术
C.语音识别技术
D.时序分析技术

12.目前深度学习技术还不能实现的应用是_____。
A.基因序列分析
B.完全模拟人的神经元认知行为
C.基于用户要求撰写宣传文案
D.人脸识别

二、多选题（ 本大题 5 道小题 ，每小题 2 分，共 10 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所有正确答案。

1.数据科学的知识结构主要由_____组成。
A.生活习惯
B.数学
C.计算机科学
D.领域专业知识

2.在分析学生身体状况时，通常需要对数据集进行_____预处理。
A.删除体重为负的数据
B.删除重复学生的数据
C.删除有缺失值的数据列
D.采用学生平均身高填充学生身高的缺失值

3.下列体现大数据特征的应用场景有_____。
A.使用 Excel 生成数据透视表
B.ChatGPT 使用互联网上过去数十年的文本、图像数据训练模型
C.ChatGPT 能够同时响应数亿用户的提问
D.统计一篇文章中词的出现频率

4.评估分类模型的性能，常用的指标包括_____。
A.Accuracy(准确率)
B.RMSE（均方根误差）
C.F1-评分
D.Precision（精确率）

5.图像识别的应用场景包括_____。
A.人脸支付
B.停车场车牌自动识别
C.卡证自动识别
D.图片亮度调整

三、程序填空题（ 本大题 4 道小题 ，每空 4 分，共 52 分）。

1. 提示

a) 题目源程序存放在"C:\KS"文件夹下，供程序调试；

b) Python科学计算库函数使用说明存放在"C:\KS"文件夹下，注意不同类库的函数存放在相应的sheet下。

调查机构对年轻人早餐支出情况进行分析，现采用随机生成数据的方法进行模拟。随机生成100位男士7天早餐消费（单位：元）数据，每人每天消费金额为[3,20)范围内的随机整数。（源程序文件fill_1.py）。

1）用numpy函数生成二维随机整数，每行表示一位消费者信息，计算每人的周平均消费金额；

2）分区间统计周均消费金额（第一类：[5,10]；第二类：>15）的人数并输出；

3）计算每人的周消费总金额，统计总额小于50元的人数。

```
#源程序文件fill_1.py
import numpy as np
import pandas as pd
#设置亚洲文字显示宽度
pd.set_option("display.unicode.east_asian_width",True)
pd.set_option("display.unicode.ambiguous_as_wide",True)
#1）为100人随机生成7天的早餐消费金额，计算每人的早餐周平均消费金额
a=np.random.【1】(3,20,(100,7))
print(a[:10])
b=a.mean(axis=1)
print(b)

#2）统计两类周平均消费金额的人数并显示
x1=b[【2】].size
print("消费金额在5至10元的人数：",x1)
x2=b[b>15].size
print("消费金额大于15元的人数：",x2)

#3）计算每人的周消费总金额，统计总额小于50元的人数。
asum=a.sum(axis=1)
y1=【3】.size
print("周消费总金额小于50元的人数：",y1)
```

2. 提示

a) 题目源程序存放在"C:\KS"文件夹下，供程序调试；

b) Python科学计算库函数使用说明存放在"C:\KS"文件夹下，注意不同类库的函数存放在相应的sheet下。

表1记录了我国到2022年7月为止部分省份新能源车与公共充电桩数据，请用这些数据进行统计分析（源程序文件fill_2.py）。具体要求如下：

1）创建保存各省份新能源车与充电桩保有量的数据对象；

2）添加一条安徽省的新能源车（41.03万辆）与充电桩（6.85万个）数据；

3）增加一列"ratio"，记录充电桩与新能源车的数量比，查询并输出比例小于10%的省份。

表1 2022年7月部分省份新能源车与公共充电桩数量

省份（province）	公共充电桩保有量（万个）（energy）	新能源汽车数量（万辆）（vehicle）
广东	33.07	199.81
江苏	10.96	97.99
浙江	10.54	133.95
山东	7.39	90.65
湖北	8.38	39.59
海南	2.35	18.61
江西	1.99	23.56
广西	2.9	43.62
云南	3.24	21.35

```
#源程序文件fill_2.py
import pandas as pd
energyArr = [["广东",33.07,199.81], ["江苏",10.96,97.99], ["浙江
",10.54,133.95], ["山东",7.39,90.65], ["湖北",8.38,39.59], ["海南
",2.35,18.61], ["江西",1.99,23.56], ["广西",2.9,43.62], ["云南",3.24,21.35]]

#设置亚洲文字显示宽度
pd.set_option("display.unicode.east_asian_width",True)
pd.set_option("display.unicode.ambiguous_as_wide",True)

#1)创建保存各省份新能源车与充电桩保有量的数据对象。
title =("province","energy","vehicle")  #对应表1的第一行3个标题
df1 = pd.DataFrame(【1】)
print(df1)

#2)添加一条安徽的新能源车与充电桩数据
de = {"province":["安徽"],"energy":[6.85],"vehicle":[41.03]}
df2 = pd.DataFrame(de)
df = pd.concat(【2】,ignore_index=True)
print(df)

#3)增加一列"ratio"，记录充电桩与新能源车的数量比，输出比例小于10%的省份
df["ratio"]=df.energy/df.vehicle
print("充电桩与新能源车的比例小于10%的省份:\n")
print(df.loc [【3】,"province"])
```

3. 提示

a) 题目源程序存放在"C:\KS"文件夹下，供程序调试；

b) Python科学计算库函数使用说明存放在"C:\KS"文件夹下，注意不同类库的函数存放在相应的sheet下。

文件gold_data.csv中存放了2012年至2022年黄金交易数据。在处理了缺失数据后，统计2022年金价的平均收盘价（源程序文件fill_3.py）。具体要求如下：

1) 读取gold_data.csv文件中的数据；

2) 判断数据中是否存在缺失值，并输出各列的缺失值情况。交易量（Volume）列缺失的值，使用前一天的交易量值填充；

3) 计算2022年金价的平均收盘价(Close/Last)。

```
#源程序文件（fill_3.py）
import pandas as pd
#设置亚洲文字显示宽度
pd.set_option("display.unicode.east_asian_width",True)
pd.set_option("display.unicode.ambiguous_as_wide",True)
#1) 读取gold_data.csv文件中的数据；
data= pd.read_csv('gold_data.csv')
print(data.head())

#2)判断数据中是否存在缺失值，并输出各列的缺失值情况；
is_NAN = data.【1】.any()
print(is_NAN)
# 交易量(Volume)列缺失的值，使用前一天的交易量值进行填充；
【2】(method='ffill', inplace=True)

#3)计算2022年金价的平均收盘价(Close/Last)；
jj=data.loc[data["Date"]>='01/01/2022','Close/Last']
print("2022年金价的平均收盘价：{:.2f}".format(【3】))
```

4. 提示

a) 题目源程序存放在"C:\KS"文件夹下，供程序调试；

b) Python科学计算库函数使用说明存放在"C:\KS"文件夹下，注意不同类库的函数存放在相应的sheet下。

文件ScenicSpots.csv记录了上海受欢迎的62个景点和所属行政区，统计分析各区的景点数（源程序文件fill_4.py）。具体要求如下：

1）从文件ScenicSpots.csv读入数据；

2）统计不同区的景点总数，并按景点总数降序排列(替换原数据)；

3）绘制柱形图（如图1所示）反映各区景点数高低情况。

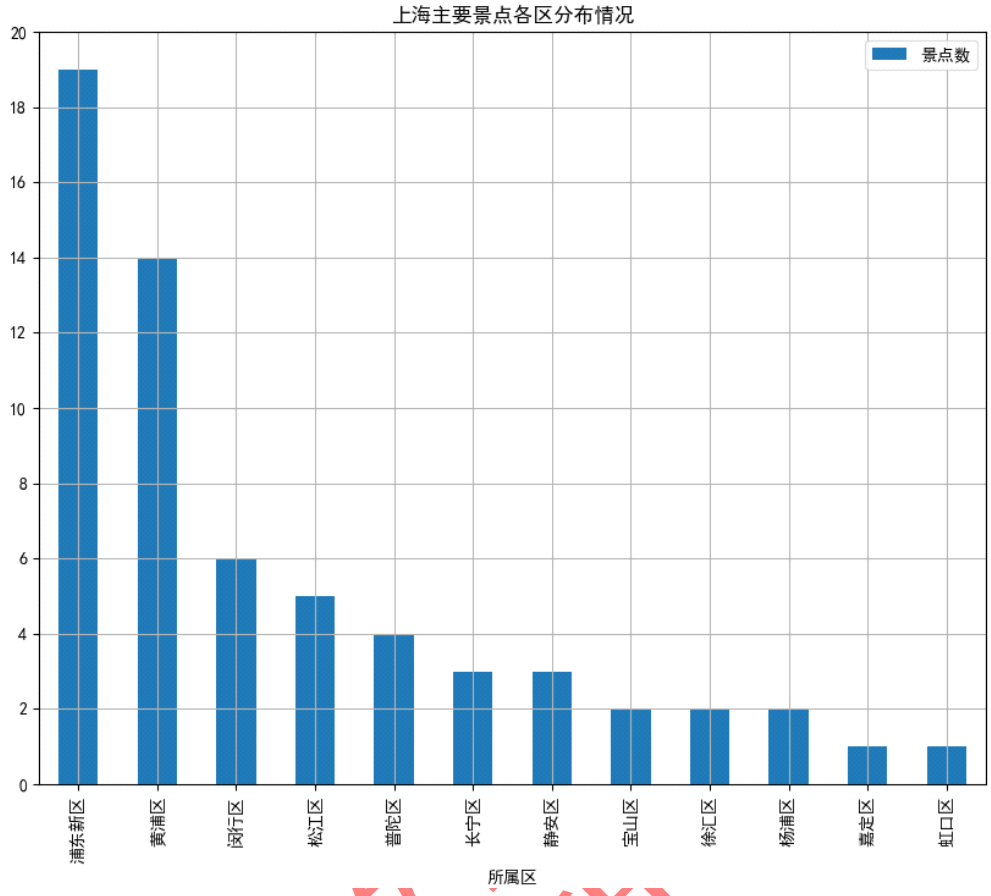

图1 各区景点数排行


```
#源程序文件（fill_4.py）
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
#设置亚洲文字显示宽度
pd.set_option("display.unicode.east_asian_width",True)
pd.set_option("display.unicode.ambiguous_as_wide",True)

#1）从文件ScenicSpots.csv读入数据
spots = pd.read_csv("ScenicSpots.csv",encoding ="gbk")
print(spots)

#2）统计不同区的景点总数，并按景点总数降序排列(替换原数据)；
tspot = spots.groupby(by = ["所属区"]).【1】
tspot.columns=['景点数']
tspot.【2】(by = ["景点数"],ascending=False,inplace=True)
print(tspot)
```

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

#3）绘制柱形图（如图1所示）反映各区景点数高低情况。
tspot.【3】
plt.title("上海主要景点各区分布情况")
plt.【4】(np.arange(0,21,2)) #将y轴刻度最大值设置为20，间隔2
plt.grid()
plt.show()
```

四、操作题

（一）、简答题（共2题，每题8分，共16分）

提示：打开C:\KS\Answer.docx文件，将简答题答案写在该文件的相应题目下并保存。

1. 高校评选奖学金需要收集学生的学业信息，请给出采集的数据项名称，说明数据的类型（连续数值/可选项/文本/图像/视频/声音 等），并给出2条以上的样例数据说明。

2. 为了帮助学生根据历年奖学金的获奖信息预测自己获奖的等级，如何建立预测模型？请描述建模的分析目标，准备数据集的方法、适用的分析模型等。

（二）、综合应用题（共10小题，48分）

提示：打开"C:\KS"文件下的程序文件"prog.py"，按照程序注释说明，编写代码实现功能要求。

文件Sleep_health.csv的数据集中记录了与睡眠和日常习惯有关的诸多特征，如性别、年龄、职业、睡眠时长、睡眠质量、身体活动水平、压力水平、BMI类别、收缩压、舒张压、心率、每日步数、以及是否有睡眠障碍等。利用提供的数据建立模型来预测睡眠障碍类型并对模型进行性能评估。具体要求如下：

1) 从文件Sleep_health.csv中读出所需的数据，将ID列为索引；(3分)
2) 对数据集进行预处理，添加'是否高血压'列，设置满足条件（'收缩压'>130 且 '舒张压'>80）的值为1，其它为0； (6分)
3) 将性别字段值["男","女"]转化为 [1,0]，将['BMI','睡眠障碍']两列数据转换为对应数值型 [0,1,2]；（6分）
4) 绘制各数值类型列的散点图矩阵，观察各因数之间的关系；(4分)
5) 统计各'职业'的['睡眠时长','睡眠质量','身体活动水平','压力水平','每日步数']的平均值；(3分)
6) 用['收缩压','舒张压','职业']以外的列建立数据集，判别睡眠障碍类型；(3分)
7) 将数据集按照合适比例分为训练集和测试集；(3分)
8) 至少选用两种以上的算法在训练集上建立分类模型，预测睡眠障碍类型；(8分)
9) 在测试集上测试各模型的预测性能；(8分)
10) 根据第9）步的运行结果，说明分类模型用于睡眠障碍类型预测的效果，比较选用模型的性能，请描述在程序文件给出的注释行中。(4分)