想要了解上海市小学生的身高,需要抽取 500 个样本,这项调查中的样本是?
正确答案: A 你的答案: 空 (错误)

从中抽取的 500 名学生的身高
上海市全部小学生的身高
从中抽取的 500 名小学生
上海市全部小学生

以下对 k-means 聚类算法解释正确的是
正确答案: C 你的答案: 空 (错误)

能自动识别类的个数,随即挑选初始点为中心点计算
能自动识别类的个数,不是随即挑选初始点为中心点计算
不能自动识别类的个数,随即挑选初始点为中心点计算
不能自动识别类的个数,不是随即挑选初始点为中心点计算

以下哪个是常见的时间序列算法模型
正确答案: C 你的答案: 空 (错误)

RSI
MACD
ARMA
KDJ

有个袋子装有 2 个红球,2 个蓝球,1 个黄球,取出球以后不再放回,请问取两次出来的球是相同颜色的概率是多少
正确答案: C 你的答案: 空 (错误)

0.3333
0.25
0.2
0.1667

65,8,50,15,37,24,()。括号中的数字是()
正确答案: B 你的答案: 空 (错误)

25
26
22
27

一组数据,均值>中位数>众数,问这组数据
正确答案: B 你的答案: 空 (错误)

左偏
右偏
钟型
对称

SQL 语言允许使用通配符进行字符串匹配的操作,其中'%'可以表示
正确答案: D 你的答案: 空 (错误)

零个字符
1 个字符
多个字符
以上都可以

关于正态分布,下列说法错误的是:
正确答案: C 你的答案: 空 (错误)

正态分布具有集中性和对称性
正态分布的均值和方差能够决定正态分布的位置和形态
正态分布的偏度为 0，峰度为 1
标准正态分布的均值为 0，方差为 1

在以下不同的场景中,使用的分析方法不正确的有
正确答案: B 你的答案: 空 (错误)

根据商家最近一年的经营及服务数据,用聚类算法判断出天猫商家在各自主营类目下所属的商
家层级
根据商家近几年的成交数据,用聚类算法拟合出用户未来一个月可能的消费金额公式
用关联规则算法分析出购买了汽车坐垫的买家,是否适合推荐汽车脚垫
根据用户最近购买的商品信息,用决策树算法识别出淘宝买家可能是男还是女

下列时间序列模型中,哪一个模型可以较好地拟合波动性的分析和预测
正确答案: D 你的答案: 空 (错误)

AR 模型
MA 模型
ARMA 模型
GARCH 模型

excel 工作簿 a 中有两列 id、age,工作簿 b 中有一列 id,需要找到工作薄 b 中 id 对应的 age,可用的函数包括
正确答案: A B 你的答案: 空 (错误)

index+match
vlookup

hlookup
find
if
like

现在有 M 个桶,每桶都有 N 个乒乓球,乒乓球的颜色有 K 种,并且假设第 i 个桶第 j 种颜色的球个数为 Cij, 比例为 Rij=Cij/N,现在要评估哪个桶的乒乓球颜色纯度最高,下列哪种算法和描述是合理的?

正确答案: C D E F 你的答案: 空 (错误)

∑(N/K-Cij)(N/K-Cij)越小越纯
-∑Cij*LOG(Rij)越小越纯
∑(1-Rij*Rij)越小越纯
∑(1-Rij)*(1-Rij)越小越纯
∑(1-Rij)^2 越小越纯
-∑Rij*LOG(Rij)越小越纯

关于相关系数,下列描述中正确的有:

正确答案: A C E 你的答案: 空 (错误)

相关系数为 0.8 时,说明两个变量之间呈正相关关系
相关系数等于 1 相较于相关系数等于-1,前者的相关性更强
相关性等于 1 相较于相关系数等于 0,前者的相关性更强
Pearson 相关系数衡量了两个定序变量之间的相关程度
Spearman 相关系数可以衡量两个定序变量之间的相关程度
相关系数为 0.2 相较于-0.8,前者的相关性更强

关于线性回归的描述,以下正确的有:

正确答案: B C E 你的答案: 空 (错误)

基本假设包括随机干扰项是均值为 0,方差为 1 的标准正态分布
基本假设包括随机干扰下是均值为 0 的同方差正态分布
在违背基本假设时,普通最小二乘法估计量不再是最佳线性无偏估计量
在违背基本假设时,模型不再可以估计
可以用 DW 检验残差是否存在序列相关性
多重共线性会使得参数估计值方差减小

下列哪些方法可以用来对高维数据进行降维:

正确答案: A B D E F 你的答案: 空 (错误)

LASSO
主成分分析法
聚类分析
小波分析法

线性判别法
拉普拉斯特征映射

查询成交表 a 中的城市 city 的成交金额大于 0 的购买人数(buyer_id)和成交金额(amt)

city buyer_id order_id amt
a 1 1 100
a 1 2 100
b 2 3 100
b 3 4 20
c 4 5 0

```
1    select buyer_id,sum(amt) as amt from a
2    where city in
3    (
4    select city from
5    (
6    select city,sum(amt) as amt from a
7    group by city
8    )t
9    where t.amt>0
10   )
```

公司要构建淘宝商家健康指数,所以要对最近 1 年内有交易的淘宝商家进行问卷调研。为不过于打搅商 家,问卷调研采取抽样的方式进行确定商家名单。怎么抽样比较好?

参考答案

可以考虑采用分层随机抽样的方式。 首先根据销售额或销售量对商家进行分层,这样可能会将商家分为高销售额(量) 商户,中销售额(量)商户,低销售额(量)商户等,然后根据这三者的比例确定 各个层次应抽取的商户数。 对抽取出来的样本,根据相应的指标,如访问量、购买量、买家评级,评论数,发 货速度等指标来综合考虑商家的健康指数。