2017阿里巴巴数据分析校园招聘笔试

21 道题，100 分，60 分钟

一、单选题（10）

1. 想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

A. 从中抽取的 500 名学生的身高
B. 上海市全部小学生的身高
C. 从中抽取的 500 名小学生
D. 上海市全部小学生

2. 以下对 k-means 聚类算法解释正确的是

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

3. 以下哪个是常见的时间序列算法模型

A. RSI
B. MACD
C. ARMA
D. KDJ

4. 有个袋子装有 2 个红球，2 个蓝球，1 个黄球，取出球之后不再放回，请问取两次出来的球是相同颜色的概率是多少

A. 0.3333
B. 0.2500
C. 0.2000
D. 0.1667

5. 65，8，50，15，37，24，（）。括号中的数字是（）

A. 25
B. 26
C. 22
D. 27

6. 一组数据，均值>中位数>众数，问这组数据

A. 左偏
B. 右偏
C. 钟型
D. 对称

7. 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

A. 零个字符

B. 1 个字符

C. 多个字符

D. 以上都是

8. 关于正态分布，下列说法错误的是

A. 正态分布具有集中性与对称性

B. 正态分布的军事与方差能够决定正态分布的位置与形态

C. 正态分布的偏度为 0，峰度为 1

D. 标准正态分布的均值为 0，方差为 1

9. 以下不同的场景中，使用分析方法不正确的有

A. 根据商家最近一年的经营与服务数据，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

B. 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

C. 用关联规则算法分析出购买汽车坐垫的买家是否适合推荐汽车脚垫

D. 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

10. 下列时间序列模型中，那个模型可以较好地拟合波动性的分析与预测

A. AR 模型

B. MA 模型

C. ARMA 模型

D. GARCH 模型

二、多选题（5）

11. Excel 工作簿 a 中有两列 id、age，工作簿 b 中有一列 id，需要找到工作簿 b 中 id 对应的 age，可用的函数包括

A. Index+match

B. Vlookup

C. Hlookup

D. Find

E. If

F. Like

12. 现在有 M 个桶，每个桶都有 N 个乒乓球，乒乓球的颜色有 K 种，并且假设第 i 个桶第 j 种颜色的球的个数为C_{ij}，比例为$R_{ij}=C_{ij}/N$，现在要求颜色纯度越高，下列哪种算法描述是合理的

A. $\sum\left(N/K-C_{ij}\right)$越小越纯

B. $-\sum C_{ij}*LOG\left(R_{ij}\right)$越小越纯

C. $\sum\left(1-R_{ij}*R_{ij}\right)$越小越纯

D. $\sum\left(1\text{-}R_{ij}\right)\left(1\text{-}R_{ij}\right)$越小越纯

E. $\sum\left(1\text{-}R_{ij}\right)^2$越小越纯

F. $-\sum R_{ij}*LOG\left(R_{ij}\right)$越小越纯

13. 关于相关系数，下列描述中正确的有：

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

14. 关于线性回归的描述，以下正确的有

A. 基本假设包括随即干扰项是均值为 0 的同方差正态分布
B. 基本假设包括随即干扰项下是均值为 0 的同方差正态分布
C. 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
D. 在违背基本假设时，模型不在可以估计
E. 可以用 DW 检验残差是否存在序列相关性
F. 多重共线性会使得参数估计值方差减少

15. 下列哪些方法可以用来对高位数据进行降维

A. LASSO
B. 主成分分析
C. 聚类分析
D. 小波分析法
E. 线性判别法
F. 拉普拉斯特征映射

三、问答题

16. 程序员 A 在某个环境中编写代码，发现这个环境中只有一个函数 rand9 能产生 1-9 这 9 个数字，请问他该如何使用这个 rand9 函数编写一个能随机产生 1-10 的 10 个数字的 rand10 函数

17. 查询成交表 a 中的城市 city 的成交额大于 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

18. 公司要构建淘宝商家健康指数，所以要对最近 1 年内交易的淘宝商家进行问卷调研。为不过于打搅商家，问卷调研采取抽样方式进行确定商家名单。怎样抽取比较好？

19. 已知 A 商家近五年每月的成交数据，请列出两种不同时间爱你序列预测模型可以用来预测商家接下来三个月的成交，并详细阐述在使用每一种方法前需要对数据进行什么预处理以及具体方法？

20. 下面数据是 2015 年 4 月 1 日至 4 月 10 日某业务的数据，请对这些数据进行分析并得出分析观点：

日期	交易量	交易笔数	客户数	新客户数	新客户交易笔数	新客户交易量
2015/4/1	594.7	16.8	13.5	1.9	2.2	65.9
2015/4/2	601.9	17.0	13.5	4.0	4.7	133.8
2015/4/3	607.2	17.4	13.8	3.7	4.4	132.8
2015/4/4	632.1	17.9	14.1	4.2	4.8	162.5
2015/4/5	685.4	19.1	15.0	5.1	6.1	192.8
2015/4/6	756.6	18.7	14.9	4.5	5.3	217.5
2015/4/7	753.4	18.2	14.5	3.6	4.1	164.7
2015/4/8	640.3	18.8	14.6	3.7	4.7	164.8
2015/4/9	1236.2	39.6	23.9	10.5	18.8	412.2
2015/4/10	664.6	19.7	15.3	3.4	4.4	145.8

注：数据单位为万

21. 你理解中的分析师是什么样的？你觉得自己应聘分析师职位的有事是什么？并说明理由

一、异常值是指什么?请列举 1 种识别连续型变量异常值的方法?

异常值(Outlier) 是指样本中的个别值，其数值明显偏离所属样本的其余观测值。在数理统计里一般是指一组观测值中与平均值的偏差超过两倍标准差的测定值。

Grubbs' test(是以 Frank E. Grubbs 命名的)，又叫 maximum normed residual test，是一种用于单变量数据集异常值识别的统计检测，它假定数据集来自正态分布的总体。

未知总体标准差 σ，在五种检验法中，优劣次序为：t 检验法、格拉布斯检验法、峰度检验法、狄克逊检验法、偏度检验法。

点评：考察的内容是统计学基础功底。

二、什么是聚类分析?聚类算法有哪几种?请选择一种详细描述其计算原理和步骤。

聚类分析(cluster analysis)是一组将研究对象分为相对同质的群组(clusters)的统计分析技术。 聚类分析也叫分类分析(classification analysis)或数值分类(numerical taxonomy)。聚类与分类的不同在于，聚类所要求划分的类是未知的。

聚类分析计算方法主要有： 层次的方法(hierarchical method)、划分方法(partitioning method)、基于密度的方法(density-based method)、基于网格的方法(grid-based method)、基于模型的方法(model-based method)等。其中，前两种算法是利用统计学定义的距离进行度量。

k-means 算法的工作过程说明如下：首先从 n 个数据对象任意选择 k 个对象作为初始聚类中心;而对于所剩下其它对象，则根据它们与这些聚类中心的相似度(距离)，分别将它们分配给与其最相似的(聚类中心所代表的)聚类;然 后再计算每个所获新聚类的聚类中心(该聚类中所有对象的均值);不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数. k 个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。

其流程如下：

(1)从 n 个数据对象任意选择 k 个对象作为初始聚类中心;

(2)根据每个聚类对象的均值(中心对象)，计算每个对象与这些中心对象的距离;并根据最小距离重新对相应对象进行划分;

(3)重新计算每个(有变化)聚类的均值(中心对象);

(4)循环(2)、(3)直到每个聚类不再发生变化为止(标准测量函数收敛)。

优点：本算法确定的 K 个划分到达平方误差最小。当聚类是密集的，且类与类之间区别明显时，效果较好。对于处理大数据集，这个算法是相对可伸缩和高效的，计算的复杂度为 O(NKt)，其中 N 是数据对象的数目，t 是迭代的次数。一般来说，K<<n，t<<n p="" 。<="">

缺点：1. K 是事先给定的，但非常难以选定;2. 初始聚类中心的选择对聚类结果有较大的影响。

点评：考察的内容是常用数据分析方法，做数据分析一定要理解数据分析算法、应用场景、使用过程、以及优缺点。

三、根据要求写出 SQL

表 A 结构如下：

Member_ID(用户的 ID，字符型)

Log_time(用户访问页面时间，日期型(只有一天的数据))

URL(访问的页面地址，字符型)

要求：提取出每个用户访问的第一个 URL(按时间最早)，形成一个新表(新表名为 B，表结构和表 A 一致)

createtable B asselectMember_ID, min(Log_time), URL from Agroup byMember_ID ;

点评：SQL 语句，简单的数据获取能力，包括表查询、关联、汇总、函数等。

另外，这个答案其实是不对的，实现有很多方法，任由大家去发挥吧。

四、销售数据分析

以下是一家 B2C 电子商务网站的一周销售数据，该网站主要用户群是办公室女性，销售额主要集中在 5 种产品上，如果你是这家公司的分析师，

a) 从数据中，你看到了什么问题?你觉得背后的原因是什么?

b) 如果你的老板要求你提出一个运营改进计划，你会怎么做?

a) 从这一周的数据可以看出，周末的销售额明显偏低。这其中的原因，可以从两个角度来看：站在消费者的角度，周末可能不用上班，因而也没有购买该产品的欲望;站在产品的角度来看，该产品不能在周末的时候引起消费者足够的注意力。

b) 针对该问题背后的两方面原因，我的运营改进计划也分两方面：一是，针对消费者周末没有购买欲望的心理，进行引导提醒消费者周末就应该准备好该产品;二是，通过该产品的一些类似于打折促销等活动来提升该产品在周末的人气和购买力。

点评：数据解读能力，获取数据是基本功，仅仅有数据获取能力是不够的，其次是对数据的解读能力。

五、用户调研

某公司针对 A、B、C 三类客户，提出了一种统一的改进计划，用于提升客户的周消费次数，需要你来制定一个事前试验方案，来支持决策，请你思考下列问题：

a) 试验需要为决策提供什么样的信息?

c) 按照上述目的，请写出你的数据抽样方法、需要采集的数据指标项，以及你选择的统计方法。

a) 试验要能证明该改进计划能显著提升 A、B、C 三类客户的周消费次数。

b) 根据三类客户的数量，采用分层比例抽样;

需要采集的数据指标项有：客户类别，改进计划前周消费次数，改进计划后周消费次数；

选用统计方法为：分别针对 A、B、C 三类客户，进行改进前和后的周消费次数的，两独立样本 T-检验(two-sample t-test)。

点评：业务理解能力和数据分析思路，这是数据分析的核心竞争力。

综上所述：一个合格的数据分析应该具备统计学基础知识、数据分析方法、数据获取、数据解读和业务理解、数据分析思想几个方面能力，即将成为数据分析师的亲们，你们准备好了吗

阿里巴巴的相关信息

1. 阿里巴巴的生日是 **1999 年。**
2. 阿里巴巴的使命是**让天下没有难做的生意。**
3. 我们的愿景是**让客户相会、工作和生活在阿里巴巴，并持续发展最少 102 年**（1999 年创办，想跨三个世纪所以定了 102 年）。
4. 双十一购物狂欢节始于 **2009 年 11 月 11 日**（天猫（淘宝商城）推出的光棍节促销）
5. 2015 年 4 月 1 日至 2016 年 3 月 21 日阿里巴巴的交易额是多少：**3 万亿**，阿里巴巴用了 13 年完成的交易金额沃尔玛用了 54 年
6. 阿里巴巴人才观的是：**人才是最好的财富、平凡的人做不平凡的事、让员工快乐的工作**
7. 阿里巴巴价值观：**客户第一、团队合作、拥抱变化、诚信、激情、敬业**
8. 阿里巴巴文化：**关乎维护小企业的利益**
9. **阿里巴巴部门**

   阿里安全、阿里健康、阿里旅行、阿里妈妈、阿里数娱、阿里通信、阿里影业、阿里云、B2B&农村淘宝、菜鸟、钉钉、高德、国际 B2C、国际 UED、集团客户体验、聚划算、蚂蚁金服、OS 事业群、商家业务、商业智能部、数据技术及产品部、数据应用部、搜索、淘宝、天猫、UC、业务平台、友盟+、中间件
10. **飞天开放平台**是阿里巴巴集团自主研发的云计算平台，负责管理数据中心 Linux 集群的物理资源，控制分布式程序运行，隐藏下层故障恢复和数据冗余等细节。