1. **数据思维谈一谈如何识别作弊用户 （至少说出 3 个方面）**[业务题]

解析：可以从以下几个方面来讨论：

（1）留存率：

有时候渠道刷量会选择在次日、7 日、30 日这些重要时间点上导入用户数据。我们会发现 APP 在次日、7 日、30 日这些关键时间点上的数据明显高于其他时间点。其实真实的用户的留存曲线是一条平滑的指数衰减曲线,如果你发现你的留存曲线存在陡升陡降的异常波动，基本上就是渠道干预了数据。

（2）用户终端：

a.关注低价设备的排名。一般伪造假用户的工作坊以低端机为主

b.关注新版本操作系统的占比。刷量工作坊一般系统更新较慢

c.关注 wifi 网络的使用情况。在高速网络的环境,无论是新增用户还是活跃用户,wifi 的使用占比都比较大,并且新增用户的 wifi 使用比例会大于启动用户的 wifi 使用比例（流量下载贵）。

d.来源地区

（3）用户行为：

a.比较用户行为数据。

如果一个 APP 做的时间比较久，访问页面、使用时长、访问间隔、使用频率等这些行为数据会趋向稳定的。不同 APP 的行为数据是有差异的。可能刷量工作室可以模拟出看似真实的用户行为，但是很难跟你的 APP 的日常数据做的完全一致。此外，一个渠道用户的使用时长、使用频率过高过低都值得怀疑。我们在平时做渠道数据分析时，可以将这些数据跟整个 APP 作比较，或者将安卓市场、应用宝这些大型应用商店的数据作为基准数据，进行比较。

b.了解新增用户、活跃用户小时时间点数据曲线。

很多刷量工作室通过批量导入设备数据或者定时启动的方式来伪造数据。这种情况下，新增和启动的曲线会出现陡增和陡降。真实用户的新增和启动是一条平滑的曲线。一般来说，用户的新增和启动会在下午 6 点之后达到高峰。而且新增相比启动的趋势会更加明显。可以将不同渠道的分时数据进行对比，找到异常。

（4）转化率分析

如果一个用户是真实的流量，他会经历点击、下载、激活、注册、直到触发目标行为的过程。我们可以将这些步骤做成漏斗模型，观察每一步的转化率。漏斗的步骤越靠后，作弊的难度越大，所获取用户对系统的价值越高，同时我们付出的用户成本也越高。运营人员需要对目标行为进行监控，在渠道推广时，考察目标行为的转化率，提高渠道作弊的边际成本。

（5）异常特征

设备号异常（频繁重置 idfa）、ip 异常（异地访问）、行为异常（突然大量点击广告、点赞）、数据包不完整等

2. **如何估算京东一日订单量？[估算]**

解析：

角度一：京东一日订单量=（中国网民数量*使用京东的比例）÷用京东购物的天数间隔

据《中国互联网络发展状况统计报告》统计显示，截至 2018 年 12 月，我国网民规模达 8.29 亿。据调查，京东市场占有比约为 16%，则假设网民中正在使用京东的比例为 16%，则有 1.326 亿人。若京东用户平均每一个月用京东购物一次，则

京东一日订单量=13260 万÷30=442 万

角度二：京东一日订单量=京东物流配送人员人数*每日可送达订单数
京东订单由京东物流负责配送，京东共有 18 万物流配送人员。根据招聘信息，一位物流配送员每天工作 10 小时，每月休息 4 天，则每天工作 8 小时 40 分钟。若配送一个订单需要 15 分钟，则一天可送约 35 单。
由此，京东一日订单量=18*35=630 万

结果分析：根据两个角度的估算，京东一日订单量约为 400~700 万。仍有一些因素可能导致误差，如京东用户使用京东购物频率尚未考证，可用过用户调查使其更为准确。

3. **如何利用 Numpy 对数列的前 n 项进行排序？[Python]**
解析：
argsort 函数返回的是 index
x[x[:n-1].argsort()]

4. **如何检验一个数据集或者时间序列是随机分布的？[Python]**
答案：画 lag plot（Correlogram：相关图），如果图上的点呈散乱分布，则为随机

5. **在 python 中如何创建包含不同类型数据的 dataframe？[Python]**
答案：利用 pandas 包的 DataFrame 函数的 serias 创建列然后用 dtype 定义类型：
Pd.DataFrame({ 'x' :pd.series(['1' ,' 2' ,' 3'],dtype=float)}

6. **K-means 里的 K 值指什么，KNN 里的 K 值指什么，二者有什么区别？[算法]**
解析：首先，大家先要知道，KNN 是监督学习算法，K-means 则是无监督学习算法。并且，这两个算法解决的是数据挖掘中的两类问题，K-Means 解决聚类问题，KNN 解决分类问题。K-means 也就是不需要事先给出分类标签，而 KNN 需要我们给出训练数据的分类标识。
最后，K 值的含义不同。K-Means 中的 K 值代表 K 类，给定 K 个类别，然后把数据向最近的 K 值靠近进行聚类分析。KNN 中的 K 值代表 K 个最接近的邻居，KNN 分类算法是分类算法中最简单的方法之一，通过测量不同特征值之间的距离来进行分类。

7. **什么是逻辑回归（logistic regression），其常见的应用场景是什么？[算法]**
逻辑斯蒂回归常指逻辑回归模型，逻辑回归是一种用于解决监督学习（Supervised Learning）问题的学习算法，进行逻辑回归用于预测参数之间组合可能输出的二分结果。
逻辑回归常见的应用场景是就是预测概率，如知道了某人的年龄，性别，血压，胆固醇水平等来预测一个人患心脏病的概率。

8. **请说出三种特征归一化（数据变换）的方法 [算法]**
答案：min-max, z-score，移动小数点位置（小数定标）
解析：说到数据归一化，小学生在这里想要先给大家讲一下数据标准化

（normalization）。数据标准化在某些比较和评价的指标处理中经常会用到，目的在于去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。

标准化的方法有很多，其中最典型的就是归一化，即将数据统一映射到[0,1]区间上。通过归一化处理，数据分析师在进行分析的时候可以提升模型收敛的速度，同时还可以提高模型的精度。

大家如果想要了解的更多，可以看看这里

https://blog.csdn.net/pipisorry/article/details/52247379

9. 在一张工资表 salary 里面，发现 2017-07 这个月的性别字段男 m 和女 f 写反了，请用一个 Updae 语句修复数据 例如表格数据是： id name gender salary month 1 A m 1000 2017-06 2 B f 1010 2017-06 [SQL]

解：

```
update salary
set gender = replace('mf', gender, '')
```

10. 四个人选举出一个骑士，统计投票数，并输出真正的骑士名字（提示，可以自己假设字段）[SQL]

解析：设表 tabe 中字段为 id，knight，vote_knight

```
select knight from table
group by vote_knight
order by count(vote_knight) limit 1
```

11. 请简述置信水平、置信区间的概念和他们之间的关系 [概统]

置信区间是我们所计算出的变量存在的范围，置信水平就是我们对于这个数值存在于我们计算出的这个范围的可信程度。
举例来讲，如果说我们有 95%的把握，真正的数值在我们所计算的范围里。
那么在这里，95%是置信水平，而计算出的范围，就是置信区间。

12~13 题 [SQL]

交易表结构为 user_id,order_id,pay_time,order_amount

12. 写 sql 查询过去一个月付款用户量（提示 用户量需去重）最高的 3 天分别是哪几天

13. 写 sql 查询做昨天每个用户最后付款的订单 ID 及金额

解析：

12) select count(distinct user_id) as c from table group by month(pay_time) order by c desc limit 3

13) select order_id, order_amount from ((select user_id, max(pay_time) as mt from table group by user_id where DATEDIFF(pay_time, NOW()) = -1 as t1) left join table as t2 where t1.user_id = t2.user_id and t1.mt == t2.pay_time)

14.一生产线生产的产品成箱包装，假设每箱平均重 50kg，标准差为 5kg，若用最大载重量为 5000kg 的汽车来承运，试用中心极限定理计算每辆车装多少箱，才能保证汽车不超载的概率大于 0.95,（设 ϕ (1.645)=0.95,其中 ϕ (x)是正标准正态分布 N(0,1)的分布函数）［概统］

解：设 $X_i\ (i=1,2,\cdots,n)$ 是装运的第*i*箱的重量，*n*是所求得箱数，由条件可以把 X_i 看作是相互独立同分布的随机变量，而总重量

$$T_n = X_1 + X_2 + \cdots + X_n$$

是独立同分布的随机变量之和。

由题意知，T_n 要小于5000千克。针对每个 X_i 来说，有 $\mu = 50, \sigma = 5$ 。根据林德伯格-莱维定理，T_n 服从正态分布*N(50n,25n)*，由此我们得出下列公式：

$$P(T \le 5000) = P\left(\frac{T_n - 50n}{5\sqrt{n}} \le \frac{5000 - 50n}{5\sqrt{n}}\right)$$

$$\approx \Phi\left(\frac{1000 - 10n}{\sqrt{n}}\right) > 0.95 = \Phi(1.645)$$

要满足以上公式，*n*需要满足

$$\frac{1000 - 10n}{\sqrt{n}} > 1.645$$

解此方程，可得*n*＜98.37，因此答案为最多装98箱。

15．男生点击率增加，女生点击率增加，总体为何减少？［概统］

解析：因为男女的点击率可能有较大差异，同时低点击率群体的占比增大。
如原来男性 20 人，点击 1 人；女性 100 人，点击 99 人，总点击率 100/120。
现在男性 100 人，点击 6 人；女性 20 人，点击 20 人，总点击率 26/120。