1. 请列出三种处理缺失值的方式

（1）进行填补，平均值、中位数可以用来填补 numerical value（数字型变量）； 众数可以用来填补 categorical value（定性变量）；

（2）根据和缺失值相关的其他列，填补对应的值，例如地址和邮编，可以根据同一个区的地址推断出邮编；

（3）当缺失值过大并且变量不那么关键时，可以直接删除那一列；当缺失值所占比例不足 5%左右，去除所在行，但是需要注意其他列的信息

2. 估算北京市一日卖出的油条数量

首先，这类估算问题会经常出现在数据分析、产品、咨询类岗位，统称为费米问题。分析这类问题可以分别从两个角度展开。根据情况，可以采用 Top-down，bottom-up 法则，即先从宏观层面，自上而下推，再由某个点横向切入，反推上去。或者也可以从需求层面和供给层面来说。然后可以对比两次推测得到的结果，如果相差不悬殊，那基本就没差啦。

然后在陈述的时候也可以需要说几句可能会出现误差的影响因素以及对结果的影响，会显得思考更加全面。具体的答案不是要求必须正确，重要的的是分析思路

这类练习题不要方，多练练思路，多看看平时的新闻报道，掌握一些基本数据 sense 就行。

油条这道题适合从供给需求两个层面上来思考问题：

角度一（需求层面）：北京市一天卖出的油条=早饭吃油条的人数*每人吃的油条的数量

北京市约有人口 2000 万人，假设 20 人中有 1 人选择早饭吃油条，则有 2000÷20=100 万人。每人每次吃 1 根油条。

因此，北京市一天卖出约 100*1=100 万根油条

角度二（供给层面）北京市一天卖出的油条=北京油条店的数目*每家店卖出的油条数目

北京市面积约 16410 平方千米，五环内面积约 735 万平方千米，若每 1 平方千米有 2 家油条店，则有 735*2=1470 家；五环外有约 15700 平方千米，若每两平方千米有 1 家油条店，则有 15700÷2=7850 家。由此，北京共有油条店约 9320 家。假设每家油条店每天卖出 100 根油条。那么，北京市一天卖出 9320*100=93.2 万根油条

结果分析：根据两个角度的估算，北京市一天可以卖出的油条数量约在 100 万左右。仍有一些因素可能导致误差，如五环内外油条店的分布密度尚待考证，可通过抽样调查使其更为精准。

3. 以下不属于无监督学习的算法为

A. 线性回归

B. PCA

C. Kmeans、

D. Word2vec

解析：当然是选 A 啦。

简单理解，监督学习就是用已知变量推导输出变量的过程，已知变量和输出变量之间有关系，比如典型监督模型线性回归，我们拿最简单的 Y=kx+b 来说，我们是知道很多自变量 X 的值来推 Y 值，这就是已知变量推导输出变量的意思，典型的监督学习模型还有回归、决策树、KNN、朴素贝叶斯（分类属于有监督学习）。而无监督学习则是数据之间需要通过计算机“学习”来探寻关系，聚类算法一般算是无监督学习，典型无监督学习有 k-means，PCA 等。

4. 下列不属于监督学习的算法是

A. PCA

B. SVM

C. 决策树

D. 逻辑回归

前面已经说到过，SVM、决策树、逻辑回归都属于监督学习算法，PCA（主成分分析）属于无监督学习。SVM 支持向量机是一种二元分类模型，决策树则通过根节点、父节点、子节点等对数据进行合理分类，逻辑回归同样也是通过概率划分数据类别，三个都是分类模型哦！

请大家一定要把这些模型类别记清楚，被面试官问到不会一脸懵。

5. 下列属于数据挖掘的重要任务目的的是（可多选）：

A. 预测

B. 关联

C. 分类

D. 聚类

解析：ABCD

以上统统都是！在数据挖掘的过程中，数据分析师就是利用各种软件、算法模型，对已有的数据进行关联、分类、聚类以及预测分析，从而在茫茫数据中找到隐藏的信息、支持商业决策哒~

6. 请简述相关系数和决定系数的概念，意义和应用。

相关系数：R

相关系数 R 是自变量 X 和因变量 Y 的协方差/标准差的乘积。

在统计学中，相关系数或皮尔逊相关系数（ Pearson correlation coefficient），是用于度量两个变量 X 和 Y 之间的相关（线性相关），其值介于-1 与 1 之间。

若值大于 0，表示正向相关，小于 0，表示负向相关，等于 0，表示不相关。

决定系数：R 平方值

其定义为：反应因变量的全部变异能通过回归关系被自变量解释的比例。R^2的在统计学中用于度量因变量的变异中可由自变量解释部分所占的比例，以此来判断统计模型的解释力。

例如当R^2为 0.8 时，表示因变量 80%变化可以用模型来解释

7. 请简述中心极限定理和大数定律分别的概念，应用以及二者之间的关系。

中心极限定理：

中心极限定理是概率论中的一组定理。中心极限定理说明，在适当的条件下，大量相互独立的随机变量的均值经适当标准化后依分布收敛于正态分布。

大数定律：

是描述相当多次数重复实验的结果的定律。根据这个定律知道，样本数量越多，则其算术平均值就有越高的概率接近期望值。

数学原理：

我们假设有 n 个独立随机变量，令他们的和为：

$$S_n = \sum_{i=1}^{n} X_i$$

大数定律（以一般的大数定律为例），它的公式为：

$$\frac{1}{n}S_n - E(X) \xrightarrow{P} 0$$

而中心极限定理的公式为：

$$\sqrt{n}(\frac{S_n}{n} - E(X)) \xrightarrow{D} N(0, \Sigma)$$

二者之间的关系：

这两个定律都是在说样本均值性质。随着 n 增大，大数定律说样本均值几乎必然等于均值。中心极限定理说，它们越来越趋近于正态分布。并且这个正态分布的方差越来越小

8. 逻辑回归和线性回归预测的值的本质区别是？

逻辑回归预测二分类问题，答案判断是 0-1 间的概率问题，常常用来做疾病预测，找出某个概率值，大于这个概率则是判定为 1-有疾病，小于这个概率则是判定为 0-没有疾病；

线性回归预测连续值，预测出一个函数以便查看数据和这个函数的拟合关系，常常用来预测房价、销售额，当知道影响房价的很多变量后，就可以通过各类变量来拟合一条连续线预测房价；

9. 请说出三种处理数据异常值的方式。

解析：异常值包括极大、极小值、错误值例如格式，符号错误等等，处理方式包括更改格式，去除 outlier 等

10. 在 SQL 里处理数据，想提取 10000 条数据的前 1%，函数怎么写？

解析：

```
SELECT …
FROM
LIMIT 100
```

其实只要选择数据的前 100 条就好啦，这是一个比较讨巧的办法！

11. 在 SQL 里处理数据，提取 10000 条数据的前 1%，并且按照降序排列，函数怎么写？

```
SELECT …
FROM
ORDER BY DESC
LIMIT 100
```

12. UV、PV、DAU、MAU、CVR、CTR 都是互联网产品常用的名词，请一一解释；

UV(Unique Visitor)：独立访客。访问您网站的一台电脑客户端为一个访客。00:00-24:00 内相同的客户端只被计算一次。

PV(Page View)：流量，即页面浏览量或点击量，用户每次刷新即被计算一次。

DAU（Daily Active User)：日活跃用户数量。通常统计一日（统计日）之内，登录或使用了某个产品的用户数（去除重复登录的用户）

MAU（Monthly Active Users）：月活跃用户数量。

CVR(Conversion Rate)：转化率。是一个衡量 CPA 广告效果的指标，简言之就是用户点击广告到成为一个有效激活或者注册甚至付费用户的转化率。

CTR（Click-Through-Rate）：点击通过率。是互联网广告常用的术语，指网络广告（图片广告

/文字广告/关键词广告/排名广告/视频广告等）的点击到达率，即该广告的实际点击次数（严格的来说，可以是到达目标页面的数量）除以广告的展现量（Show content）。

13. 下列描述特征工程正确的答案有（可多选）：

A. 主成分分析提取特征

B. 建立好的特征需要很强的业务知识

C. 好的特征比算法还要关键

D. 独热编码是一个重要的特征工程步骤

解析：有这么一句话大家要好好了解：“数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已”；意思大致就是算法模型只是帮助你完成使用数据的一个工具，而数据本身和特征建立会直接影响到你用的方法的效果上限。特征工程其实就是从原始数据中提取特征为算法模型使用。特征处理包括数据预处理、特征选择、降维等等，好的特征工程不仅考量算法模型，还考量对于业务的理解深刻度。

举个例子，假如你们在淘宝做数据分析的工作，在做淘宝的商品推荐排序这件事情，那么在特征上你要想，哪些是强相关的因素？哪些是弱相关？可能产品销量、价格、用户消费水平、年龄、位置都是强相关因素，但用户 ID 则是弱相关甚至无相关。我们确定强相关的因素之后，需要把所有特征都数值化，本来就是数字型变量的不用管，分类变量可以数值化，比如把积分等级、位置都数值化，更方便在模型中的操作。然后与用户行为的一些信息做整合，可以得出相关性的统计意义的结论。这也就是特征工程的意义。

当做完特征选择之后，还可能面临特征数量过大（简单理解为列数和行数太多）会影响模型训练，因此又需要降低数据维度，也就是运用主成分分析法来进行降维（关于 PCA 我们会在后面的学习中讲到）。

独热编码其实是用于数据预处理的一个方法，数据里会有很多分类变量比如国籍、性别等不能直接用数值表示的变量。独热编码可以简单理解为把这类变量转化为连续、有序的数值变量，每一个编码有且只对一个变量。过程和例子小学生不在这里细讲啦，给大家附一个网址可以看看！ https://www.imooc.com/article/35900

14. 关系数据模型的三要素是什么

数据模型的三要素：数据结构、数据操作、数据约束

数据模型所描述的内容包括三个部分：数据结构、数据操作、数据约束。

1）数据结构:数据模型中的数据结构主要描述数据的类型、内容、性质以及数据间的联系等。数据结构是数据模型的基础，数据操作和约束都建立在数据结构上。不同的数据结构具有不同的操作和约束。

2）数据操作:数据模型中数据操作主要描述在相应的数据结构上的操作类型和操作方式。

3）数据约束：数据模型中的数据约束主要描述数据结构内数据间的语法、词义联系、他们之间的制约和依存关系，以及数据动态变化的规则，

15. （今日算法）请简述线性回归模型的原理，应用和优缺点。

见群内补充讲解~