## 如何将类别型变量引入线性回归，提出至少1种方案？

1. Replace 直接替换：

直接额外定义某一类型变量为某一数字，如男女编码为1，0

1. Map() 替换：

给出一定的映射关系，类似于给出字典，根据字典查找出分类所对应的代码，没有映射关系的返回NaN。与第一种方法基本一致，好处可能在于可以对类别给出更完整定义，并非随机直接定义。

1. LabelEncoding 标签编码

LabelEncoding根据原各唯一取值的先后顺序进行排序后为其转换为对应次序的数值。对于空间有序的分类存在优先级分类，使用这种数据数据的处理分类更加适合，但是对于随即无序数据而言，引入了无关的优先级排序，会对数据分析产生不良后果。

1. OneHot Encoding独热编码

是对没有优先级的类别型数据进行处理的一种方法。生成新的n个只包含0和1的向量，表示原变量的n个唯一取值，不同的类型存储在垂直的空间，生成一个更庞大的稀疏矩阵，对数据存储空间要求大，但不会对无需数据引入有序性。

## 自变量进行标准化会对模型带来哪些影响（尝试从梯度、回归系数等角度）？

进行标准化后，所有数据的均值变为0，方差变为1，可以用于预防范围较大的特征对预测结果进行较大的影响，

**梯度**：

将分散的数据压缩到0值附近，可以使其梯度变大，结合梯度下降公式可以得知梯度变大可以加快模型参数的收敛速度，更快达到平衡值。

**回归系数**：

在回归分析中需要测量某一变量的重要程度。使用标准化过后的独立变量来训练模型，然后计算它们对应的标准化系数的绝对值差就能得出结论。然而，如果独立变量是未经标准化的，那比较它们的系数将毫无意义。

Lasso回归和岭回归对各变量对应的系数进行惩罚。变量的范围将会影响到他们对应系数受到什么程度的惩罚。因为方差大的变量对应的系数很小，因此它们会受到较小的惩罚。因此，在使用上面的两个回归之前需要进行标准化