**数据说明：**

随着大数据、人工智能等技术的发展和数字化转型提速，“数字人才”需求旺盛。leetcode是一个代码在线评测平台，提供多种考察不同算法和数据结构的编程题目，帮助用户训练算法思维和编程能力，尤其在求职者中备受欢迎。

本次作业数据“lc\_data.csv”从leetcode题库中抓取除会员专享题目外所有题目信息共2635条，分析题目难度的影响因素，因变量为是否为简单题，自变量共12个变量，可分为题目内容、题目标签和公司标签共三个类别。字段说明如下：



**分析任务：**

1. 读入数据并完成简单的数据清洗，创建EASY变量，表示是否为简单题，取值1/0分别表示简单题/非简单题；
2. 对分类变量设置基准组，其中EASY, quesclass, example的基准组分别设置为0、题库、1；
3. 对数据中的变量进行描述性分析，**并进行解读**，**包括但不一定限于**：
4. leetcode的题目难度标签有EASY/MEDIUM/HARD，请绘制三种不同题目难度占比分布直方图；
5. 题目要求的文本内容是题目最基本和核心的信息，一般来说题干越复杂，理解难度越大。请绘制不同题目难度的题目要求文本长度箱线图
6. 绘制不同题目难度的题目给出的示例数量占比柱状图；
7. 绘制不同公司的题目数量柱状图；
8. 对不同的分类标签数量，绘制简单题所占的比例直方图；
9. 建模数据预处理：由于样本中简单题共726条（占比27.55%），存在明显的类别不平衡性，可能导致分类模型的偏向。因此，请根据EASY变量的取值，对非简单样本（EASY=0）随机抽取726条数据（和简单样本数据量相同），使得数据中该变量的分布更为均衡。对平衡样本进行训练集与测试集的分割，样本比例7:3，随机树种子为2023；

提示：可以采用**ROSE包**的**ovun.sample**函数进行重采样，参数设置中，使用EASY作为因变量，方法选择“under sampling”，通过参数N的设定来设置抽取样本的数量，设置随机数种子为2023，也可以自定义采样函数完成。采样后的样本数量约为1452条；

1. 建模：使用训练集数据，EASY作为因变量，完成以下建模任务：

【决策树】

* 1. 建立CART决策树模型，找出具有较高的相对重要性的变量；

提示：在R中，可以采用**rpart包中的rpart()函数**。其中设置参数method选择决策树类型，"class"和"anova"分别对应分类树和回归树；参数parms用于分类树参数设置，parms=list(split,prior,loss)，其中split中"gini"为默认CART，"information"为ID3；参数control中可具体设置树的最大深度等形状。

* 1. 对决策树结果可视化，并对结果进行解读；

提示：可采用rpart.plot可视化rpart树的分类过程，图中每个节点的颜色及内部数字展示了节点样本分布情况

【随机森林】

* 1. 建立随机森林模型，查看变量重要性，绘制变量特征重要性图；

提示：在R中可使用randomForest包构建随机森林模型。随机森林模型包含众多参数，参数调优是提高模型表现的关键环节；在randomForest中可用importance查看变量重要性，也可用varImpPlot绘制特征重要性图

* 1. 对模型结果及变量重要性进行解读；

【boosting】

* 1. 建立adaboost模型，并对变量的重要性进行解读；

提示：在R中可使用adabag包中的boosting()函数构建adaboost模型

【综合比较】

* 1. 使用以上三个模型的建模结果，对测试集数据进行预测，绘制ROC曲线并计算AUC值，对比三个模型的预测精度，并解读结果

提示：采用pROC包中的roc()函数和ggroc()函数，绘制roc曲线并计算auc值。