**数据说明：**

大型开放式网络课堂，即MOOC（Massive open online courses）是最近几年迅速升温的一个领域，众多以在线课程为核心的互联网公司纷纷涌现或取得蓬勃发展。那么在线网络课程的收入主要受到哪些因素的影响呢？是课程内容，人数规模，还是课时安排？为了探索这一问题，我们选取了国内某在线网络课程平台在2016年8月16号这一天的全部10222个课程的数据进行分析。数据包含两个csv文件：“课程列表.csv”，“课程信息.csv”。

课程列表字段解释如下：

课程小类：s.ctg; 课程大类：b.ctg; 网址：url; 名称：course.name; 老师名字：teacher; 人数：size; 标价：real.price; 原价：orig.price; 收入：revenue;

课程信息字段解释如下：

课程网址：url; 评分：score; 适合人群：people; 课程简介：intro; 课时安排：detail;

**分析任务：**

1. 读入课程列表和课程信息两个数据集，分别命名为a和b，并用summary函数展示数据。
2. 将a、b两个数据集按照url进行合并，合并后的数据集命名为a，在合并时首先注意检查两个数据集中是否有重复观测，如果有，需要先删除重复观测。
3. 首先对课程名称进行分词，找出前20个高频词，展示一个数据框，其中第一列代表词根，第二列代表词频
4. 在任务3的基础上，统计课程名称中包含了前20个高频词中的几个并作为一个新变量，用freq.count表示并用summary函数展示freq.count变量的分布情况
5. 提取【课程信息】这个变量里的课时信息，提取之后，计算每个课程总共有多少课时，将其命名为一个新变量keshi，用summary函数展示该变量
6. 将任务5新生成的keshi变量拼接到a数据集中，筛选并保留revenue大于0的样本用于后续建模任务
7. **设置种子2022**，按照7:3划分训练集和测试集，以对数收入做因变量Y，以real.price,score,freq.count,keshi为自变量X（对real.price，keshi做必要的对数变换）构造对数线性模型并计算模型在测试集上的均方误差
8. 使用与任务7相同的训练集和测试集，考虑给出的所有自变量以及任务4和任务5中构造的变量，利用CART回归树建模并剪枝，可视化该决策树，最后计算该决策树测试集上的均方误差
9. 使用与任务7相同的训练集和测试集，考虑给出的所有自变量以及任务4和任务5中构造的变量，使用Adaboost算法建模并计算该决策树测试集上的均方误差
10. 使用与任务7相同的训练集和测试集，考虑给出的所有自变量以及任务4和任务5中构造的变量，使用随机森林建模并输出变量重要性，最后计算该决策树测试集上的均方误差
11. 比较任务7至任务10中四个模型的结果，比较这四个模型各自的运行的时间、在测试集上的均方误差。