**数据说明：**

互联网的出现和扩张，改变了人们的生活。其中之一，就是为术业有专攻的人们提供了发光发热的新方式。本案例将探索这股知识网红热潮中的一个分支，如何通过一对一聊天的方式来建立自己的职业声誉。我们选取了知识共享网站在行上的行家约聊数据，探索了行家受欢迎程度的分布特征和影响因素，并根据数据分析的结果为想要进入这个领域的专业人士提供了自我包装的建议。数据共分为四个文件，“D-cate.csv”为话题分类，“D-mentor.csv”为行家属性，“D-topic.csv”为话题属性、“D-comment.csv”为用户评价信息。由于变量较多，具体变量含义请参照“REDAME.txt”。

**分析任务：**

1. 读入数据：
   1. 读入所有话题的分类属性D-cate.csv，命名为cate
   2. 读入行家属性D-mentor.csv，命名为mentor
   3. 读入话题属性D-topic.csv，命名为topic
   4. 读入用户对话题做出的评价D-comment.csv，命名为comment
2. 数据处理：
   1. 将以下变量转化为因子型

对mentor表中的行家接受率arating，设置高接受率、中接受率、其他共三个水平；对mentor表中的行家回应邀约时间react\_tm，设置半天内回应、1天内回应、2天内回应、其他共四个水平；对topic表中的话题约见城市topic\_city，设置北京、上海、深圳、广州、杭州、成都、武汉、西安、宁波、其他共10个水平。

* 1. 构造变量：

在topic表中增加9个0-1变量：d1-d9，分别对应心理、投资理财、职场发展、教育学习、创业和融资、生活服务、互联网+、兴业经验、其他共9个话题；在topic中增加0-1变量：is\_ord，话题是否成交（topic\_ordcnt>0）；在topic中增加0-1变量：is\_fst，话题是否发生在北上深；topic中增加话题价格带：price\_b，分为<=300、(300,600]、(600,1000)、1000+四个水平；对price\_b因子化，并设定各取值的顺序；topic表中增加话题描述长度：len\_desc，话题描述字符数，划分为(0,200]、(200,400]、(400,600]、(600,800]、(800,1000]、1000+六个水平；topic表中增加行家是否回复用户评价：is\_rpl，行家是否回复用户评价（聚合comment表中is\_reply得到reply\_cnt，reply\_cnt>0）。

3. 绘制导师约见人数分布直方图（可以考虑对mentor$mentor\_ordcnt做对数变换），并进行描述分析

4. 分领域（心理、投资理财等共9个领域）绘制行家约见人数（topic\_ordcnt）的箱线图，按中位数从左到右，从高到低的顺序排列，并进行描述分析

5. 绘制不同邀约接受率（arating）、不同响应时长（react\_tm）、是否回复评价（is\_rp）和话题约见人数（topic\_ordcnt）的箱线图，并进行描述分析

6. 根据comment表中信息，绘制各行业用户数量柱状图并进行描述分析

7. 使用is\_ord作为因变量Y，自变量选择话题价格、话题价格带、想见行家用户数、行家约见人数、话题持续时长、话题所在城市、行家回复邀约的时长、行家接受约见的比率、话题长度、话题是否在北上广以及所属行业进行建模。使用AIC进行模型选择，绘制ROC曲线并计算AUC。