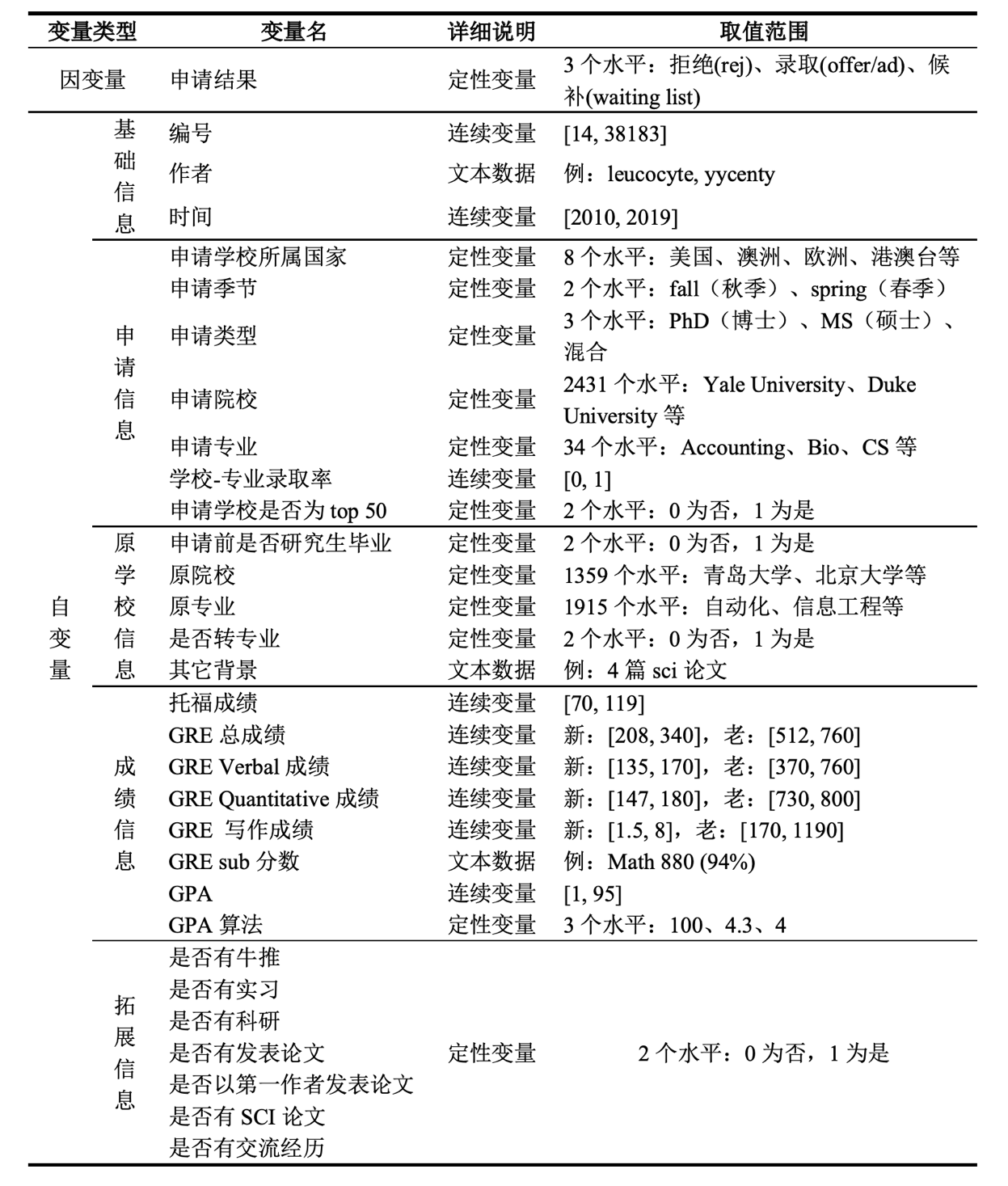
**数据说明：**

本案例使用一个留学申请数据集（Data\_Cleaning.csv），该数据来自某留学申请论坛的录取汇报结果，包含了15908条申请者的申请学校及录取与否信息。辅助的数据材料包括QS大学排名前一百名单中的美国大学（QS大学排名前百（美国）.txt）名单和美国大学缩写汇总名单（美国大学缩写汇总.txt）

留学申请数据集详细的变量含义见下表（中英文变量对照表请参考文件夹中的txt文档）。



**分析任务：**

1. 读入数据并完成简单的数据清洗，仅保留**录取与拒绝**两种结果所对应的数据记录，主要包含以下几个方面：
2. 由于数据存在错误，我们需要对申请学校这一变量进行简单修正：将学校名”Texas A”、“M University”替换”Texas A&M University”，将”Washington University in St”、” Louis”替换为”Washington University in St. Louis”。
3. 为了统一申请学校名称，我们要将学校缩写替换为全称，得到的新变量College\_apply\_new 并入原数据（注意：不考虑缩写字母的大小写差异，需要借助文件美国大学缩写汇总.txt）。
4. 读入“QS大学排名前百（美国）.txt”，数据集前19所学校是世界前五十名的美国名校，据此将所有申请学校划分为”Top50”和”Others”两类，得到因子型的新变量CollegeRankTop50并入原数据。
5. 对成绩进行离散化处理，具体地，将GPA划分为”<=3.4”, “3.4-3.55”, “3.55-3.7”, “>3.7”四类，得到新变量gpa\_dis并入原数据；将变量toefl转化为数值变量，随后划分为”<=98”, “98-102”, “102-106”, “>106”四个分数段，得到因子型的新变量toefl\_dis并入数据；
6. 删除含有缺失值的数据。
7. 对数据中的变量进行描述性分析，包括但不一定限于：

1）Top 10热门学校申请人数柱状图（请在图中分别展示不同录取情况的人数）；

2）成绩（任选GPA、托福、GRE其一）与录取结果箱线图（请根据学校的世界排名，在图中区分Top 50的学校与其他学校）；

1. 将数据集按照4:1的比例划分为训练集与测试集，划分时请固定随机数种子为2023。使用录取结果作为因变量，使用如下自变量，在训练集上拟合逻辑回归模型，并使用AIC准则进行变量筛选，解读筛选后的模型系数。

自变量包含：申请季节、申请类型、申请学校是否为Top50、拓展信息中的所有变量、GPA成绩区间、托福成绩区间。

1. 对第3问得到的筛选后的模型对测试集的申请结果进行预测，并利用R包pROC绘制出ROC曲线图，根据曲线对模型进行评价。
2. 给定预测概率的阈值为0.6，请计算对应的精准率（precision）和召回率（recall），并计算和展示混淆矩阵。
3. 绘制成本收益曲线，对该曲线内容进行解读。
4. 开放性作答：除了第3问给出的自变量外，你还可以通过数据集构造哪些自变量，从而提升模型的预测效果？请分别展示1）构造新变量的依据（可以通过描述性分析进行展示）；2）引入新变量后的模型拟合结果；3）以及对同样的测试集进行预测的效果评估（评估标准可以选择第4-6问中的任意两项标准）。